View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by HAL-UNICE

Mixing Least-Squares Estimators when the Variance is Unknown

Christophe Giraud

▶ **To cite this version:**

Christophe Giraud. Mixing Least-Squares Estimators when the Variance is Unknown. 30 pages. 2007. <hal-00184869>

HAL Id: hal-00184869

https://hal.archives-ouvertes.fr/hal-00184869

Submitted on 2 Nov 2007

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

MIXING LEAST-SQUARES ESTIMATORS WHEN THE VARIANCE IS UNKNOWN

CHRISTOPHE GIRAUD

Abstract. We propose a procedure to handle the problem of Gaussian regression when the variance is unknown. We mix least-squares estimators from various models according to a procedure inspired by that of Leung and Barron [17]. We show that in some cases the resulting estimator is a simple shrinkage estimator. We then apply this procedure in various statistical settings such as linear regression or adaptive estimation in Besov spaces. Our results provide non-asymptotic risk bounds for the Euclidean risk of the estimator.

hal-00184869, version 1 - 2 Nov 2007

1. Introduction

We consider the regression framework, where we have noisy observations

$$Y_i = \mu_i + \sigma\varepsilon_i, \quad i = 1, \dots, n \tag{1}$$

of an unknown vector $\mu = (\mu_1, \dots, \mu_n)' \in \mathbb{R}^n$. We assume that the ε_i's are i.i.d standard Gaussian random variables and that the noise level $\sigma > 0$ is unknown. Our aim is to estimate μ.

In this direction, we introduce a finite collection $\{\mathcal{S}_m,\ m \in \mathcal{M}\}$ of linear spaces of $\mathbb{R}^n$, which we call henceforth models. To each model $\mathcal{S}_m$, we associate the least-squares estimator $\hat{\mu}_m = \Pi_{\mathcal{S}_m} Y$ of μ on $\mathcal{S}_m$, where $\Pi_{\mathcal{S}_m}$ denotes the orthogonal projector onto $\mathcal{S}_m$. The L^2-risk of the estimator $\hat{\mu}_m$ with respect to the Euclidean norm $\|\cdot\|$ on $\mathbb{R}^n$ is

$$\mathbb{E}\left[\|\mu - \hat{\mu}_m\|^2\right] = \|\mu - \Pi_{\mathcal{S}_m}\mu\|^2 + \dim(\mathcal{S}_m)\sigma^2. \tag{2}$$

Two strategies have emerged to handle the problem of the choice of an estimator of μ in this setting. One strategy is to select a model $\mathcal{S}_{\hat{m}}$ with a data driven criterion and use $\hat{\mu}_{\hat{m}}$ to estimate μ. In the favorable cases, the risk of this estimator is of order the minimum over $\mathcal{M}$ of the risks (2). Model selection procedures have received a lot of attention in the literature, starting from the pioneer work of Akaike [1] and Mallows [18]. It is beyond the scope of this paper to make an historical review of the topic. We simply mention in the Gaussian setting the papers of Birgé and Massart [7, 8] (influenced by Barron and Cover [5] and Barron, Birgé and Massart [4]) which give non-asymptotic risk bounds for a selection criterion generalizing Mallows'C_p.

Date: First draft 03/02/2007, Revision 02/11/2007.

2000 *Mathematics Subject Classification.* 62G08.

Key words and phrases. Gibbs mixture - shrinkage estimator - oracle inequalities - adaptive estimation - linear regression.

An alternative to model selection is mixing. One estimates μ by a convex (or linear) combination of the $\hat{\mu}_m$s

$$\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}_m, \tag{3}$$

with weights w_m which are $\sigma(Y)$-measurable random variables. This strategy is not suitable when the goal is to select a single model $\mathcal{S}_{\hat{m}}$, nevertheless it enjoys the nice property that $\hat{\mu}$ may perform better than the best of the $\hat{\mu}_m$s. Various choices of weights w_m have been proposed, from an information theoretic or Bayesian perspective. Risk bounds have been provided by Catoni [11], Yang [25, 28], Tsybakov [23] and Bunea *et al.* [9] for regression on a random design and by Barron [3], Catoni [10] and Yang [26] for density estimation. For the Gaussian regression framework we consider here, Leung and Barron [17] propose a mixing procedure for which they derive a precise non-asymptotic risk bound. When the collection of models is not too complex, this bound shows that the risk of their estimator $\hat{\mu}$ is close to the minimum over $\mathcal{M}$ of the risks (2). Another nice feature of their mixing procedure is that both the weights w_m and the estimators $\hat{\mu}_m$ are build on the same data set, which enable to handle cases where the law of the data set is not exchangeable. Unfortunately, their choice of weights w_m depends on the variance σ^2, which is usually unknown.

In the present paper, we consider the more practical situation where the variance σ^2 is unknown. Our mixing strategy is akin to that of Leung and Barron [17], but is not depending on the variance σ^2. In addition, we show that both our estimator and the estimator of Leung and Barron are simple shrinkage estimators in some cases. From a theoretical point of view, we relate our weights w_m to a Gibbs measure on $\mathcal{M}$ and derive a sharp risk bound for the estimator $\hat{\mu}$. Roughly, this bound says that the risk of $\hat{\mu}$ is close to the minimum over $\mathcal{M}$ of the risks (2) in the favorable cases. We then discuss the choice of the collection of models $\{\mathcal{S}_m, m \in \mathcal{M}\}$ in various situations. Among others, we produce an estimation procedure which is adaptive over a large class of Besov balls.

Before presenting our mixing procedure, we briefly recall that of Leung and Barron [17]. Assuming that the variance σ^2 is known, they use the weights

$$w_m = \frac{\pi_m}{\mathcal{Z}} \exp\left(-\beta\left[\|Y - \hat{\mu}_m\|^2/\sigma^2 + 2\dim(\mathcal{S}_m) - n\right]\right), \quad m \in \mathcal{M} \tag{4}$$

where $\{\pi_m,\ m \in \mathcal{M}\}$ is a given prior distribution on $\mathcal{M}$ and $\mathcal{Z}$ normalizes the sum of the w_ms to one. These weights have a Bayesian flavor. Indeed, they appear with $\beta = 1/2$ in Hartigan [15] which considers the Bayes procedure with the following (improper) prior distribution: pick an m in $\mathcal{M}$ according to π_m and then sample μ "uniformly" on $\mathcal{S}_m$. Nevertheless, in Leung and Barron [17] the role of the prior distribution $\{\pi_m,\ m \in \mathcal{M}\}$ is to favor models with low complexity. Therefore, the choice of π_m is driven by the complexity of the model $\mathcal{S}_m$ rather than from a prior knowledge on μ. In this sense their approach differs from the classical Bayesian point of view. Note that the term $\|Y - \hat{\mu}_m\|^2/\sigma^2 + 2\dim(\mathcal{S}_m) - n$ appearing in the weights (4) is an unbiased estimator of the risk (2) rescaled by σ^2. The size of the weight w_m then depends on the difference between this estimator of the risk (2) and $-\log(\pi_m)$, which can be thought as a complexity-driven penalty (in the spirit of Barron and Cover [5] or Barron *et al.* [4]). The parameter β tunes the balance between this two terms. For $\beta \leq 1/4$, Theorem 5 in [17] provides a sharp risk bound for the procedure.

The rest of the paper is organized as follows. We present our mixing strategy in the next section and express in some cases the resulting estimator $\hat{\mu}$ as a shrinkage estimator. In Section 3, we state non-asymptotic risk bounds for the procedure and discuss the choice of the tuning parameters. Finally, we propose in Section 4 some weighting strategies for linear regression or for adaptive regression over Besov balls. Section 5 is devoted to a numerical illustration and Section 6 to the proofs. Additional results are given in the Appendix.

We end this section with some notations we shall use along this paper. We write $|m|$ for the cardinality of a finite set m, and $< x, y >$ for the inner product of two vectors x and y in $\mathbb{R}^n$. To any real number x, we denote by $(x)_+$ its positive part and by $\lfloor x \rfloor$ its integer part.

2. The estimation procedure

We assume henceforth that $n \geq 3$.

2.1. **The estimator.** We start with a finite collection of models $\{\mathcal{S}_m,\ m \in \mathcal{M}\}$ and to each model $\mathcal{S}_m$ we associate the least-squares estimator $\hat{\mu}_m = \Pi_{\mathcal{S}_m} Y$ of μ on $\mathcal{S}_m$. We also introduce a probability distribution $\{\pi_m,\ m \in \mathcal{M}\}$ on $\mathcal{M}$, which is meant to take into account the complexity of the family and favor models with low dimension. For example, if the collection $\{\mathcal{S}_m,\ m \in \mathcal{M}\}$ has (at most) e^{ad} models per dimension d, we suggest to choose $\pi_m \propto e^{(a+1/2)\dim(\mathcal{S}_m)}$, see the example at the end of Section 3.1. As mentioned before, the quantity $-\log(\pi_m)$ can be interpreted as a complexity-driven penalty associated to the model $\mathcal{S}_m$ (in the sense of Barron *et al.* [4]). The performance of our estimation procedure depends strongly on the choice of the collection of models $\{\mathcal{S}_m,\ m \in \mathcal{M}\}$ and the probability distribution $\{\pi_m,\ m \in \mathcal{M}\}$. We detail in Section 4 some suitable choices of these families for linear regression and estimation of BV or Besov functions.

Hereafter, we assume that there exists a linear space $\mathcal{S}_* \subset \mathbb{R}^n$ of dimension $d_* < n$, such that $\mathcal{S}_m \subset \mathcal{S}_*$ for all $m \in \mathcal{M}$. We will take advantage of this situation and estimate the variance of the noise by

$$\hat{\sigma}^2 = \frac{\|Y - \Pi_{\mathcal{S}_*} Y\|^2}{N_*} \tag{5}$$

where $N_* = n - d_*$. We emphasize that we do not assume that $\mu \in \mathcal{S}_*$ and the estimator $\hat{\sigma}^2$ is (positively) biased in general. It turns out that our estimation procedure does not need a precise estimation of the variance σ^2 and the choice (5) gives good results. In practice, we may replace the residual estimator $\hat{\sigma}^2$ by a difference-based estimator (Rice [20], Hall *et al.* [14], Munk *et al.* [19], Tong and Wang [22], Wang *et al.* [29], etc) or by any non-parametric estimator (e.g. Lenth [16]), but we are not able to prove any bound similar to (12) or (13) when using one of these estimators.

Finally, we associate to the collection of models $\{\mathcal{S}_m,\ m \in \mathcal{M}\}$, a collection $\{L_m,\ m \in \mathcal{M}\}$ of non-negative weights. We recommend to set $L_m = \dim(\mathcal{S}_m)/2$, but any (sharp) upper bound

of this quantity may also be appropriate, see the discussion after Theorem 1. Then, for a given positive constant β we define the estimator $\hat{\mu}$ by

$$\hat{\mu} = \sum_{m \in \mathcal{M}} w_m \hat{\mu}_m \quad \text{with} \quad w_m = \frac{\pi_m}{\mathcal{Z}} \exp\left(\beta \frac{\|\hat{\mu}_m\|^2}{\hat{\sigma}^2} - L_m \right), \tag{6}$$

where $\mathcal{Z}$ is a constant that normalizes the sum of the w_ms to one. An alternative formula for w_m is $w_m = \pi_m \exp\left(-\beta \|\Pi_{\mathcal{S}_*} Y - \hat{\mu}_m\|^2/\hat{\sigma}^2 - L_m\right)/\mathcal{Z}'$ with $\mathcal{Z}' = e^{-\beta \|\Pi_{\mathcal{S}_*} Y\|^2/\hat{\sigma}^2} \mathcal{Z}$. We can interpret the term $\|\Pi_{\mathcal{S}_*} Y - \hat{\mu}_m\|^2/\hat{\sigma}^2 + L_m/\beta$ appearing in the exponential as a (biased) estimate of the risk (2) rescaled by σ^2. As in (4), the balance in the weight w_m between this estimate of the risk and the penalty $-\log(\pi_m)$ is tuned by β. We refer to the discussion after Theorem 1 for the choice of this parameter. We mention that the weights $\{w_m,\, m \in \mathcal{M}\}$ can be viewed as a Gibbs measure on $\mathcal{M}$ and we will use this property to assess the performance of the procedure.

We emphasize in the next section, that $\hat{\mu}$ is a simple shrinkage estimator in some cases.

2.2. **A simple shrinkage estimator.** In this section, we focus on the case where $\mathcal{M}$ consists of all the subsets of $\{1, \dots, p\}$, for some $p < n$ and $\mathcal{S}_m = \text{span}\{v_j,\ j \in m\}$ with $\{v_1, \dots, v_p\}$ an orthonormal family of vectors in $\mathbb{R}^n$. We use the convention $\mathcal{S}_\emptyset = \{0\}$. An example of such a setting is given in Section 4.2, see also the numerical illustration Section 5. Note that $\mathcal{S}_*$ corresponds here to $\mathcal{S}_{\{1,\dots,p\}}$ and $d_* = p$.

To favor models with small dimensions, we choose the probability distribution

$$\pi_m = \left(1 + \frac{1}{p^\alpha}\right)^{-p} p^{-\alpha |m|}, \quad m \in \mathcal{M}, \tag{7}$$

with $\alpha > 0$. We also set $L_m = b|m|$ for some $b \geq 0$.

Proposition 1. *Under the above assumptions, we have the following expression for* $\hat{\mu}$

$$\hat{\mu} = \sum_{j=1}^{p} (c_j Z_j) v_j, \quad \textit{with} \quad Z_j = < Y, v_j > \textit{and} \quad c_j = \frac{\exp\left(\beta Z_j^2/\hat{\sigma}^2\right)}{p^\alpha \exp(b) + \exp\left(\beta Z_j^2/\hat{\sigma}^2\right)}. \tag{8}$$

The proof of this proposition is postponed to Section 6.1. The main interest of Formula (8) is to allow a fast computation of $\hat{\mu}$. Indeed, we only need to compute the p coefficients c_j instead of the 2^p weights w_m of formula (6).

The coefficients c_j are shrinkage coefficients taking values in $[0, 1]$. They are close to one when Z_j is large and close to zero when Z_j is small. The transition from 0 to 1 occurs when $Z_j^2 \approx \beta^{-1}(b + \alpha \log p)\hat{\sigma}^2$. The choice of the tuning parameters α, β and b will be discussed in Section 3.2.

Remark 1. Other choices are possible for $\{\pi_m,\ m\in\mathcal{M}\}$ and they lead to different c_js. Let us mention the choice $\pi_m = \Big((p+1)\binom{p}{|m|}\Big)^{-1}$ for which the c_js are given by

$$c_j = \frac{\int_0^1 q\prod_{k\neq j}\left[q+(1-q)\exp\left(-\beta Z_k^2/\hat\sigma^2+b\right)\right]dq}{\int_0^1 \prod_{k=1}^p\left[q+(1-q)\exp\left(-\beta Z_k^2/\hat\sigma^2+b\right)\right]dq}, \qquad \text{for } j=1,\dots,p.$$

This formula can be derived from the Appendix of Leung and Barron [17].

Remark 2. When the variance is known, we can give a formula similar to (8) for the estimator of Leung and Barron [17]. Let us consider the same setting, with $p\le n$. Then, when the distribution $\{\pi_m,\ m\in\mathcal{M}\}$ is given by (7), the estimator (3) with weights w_m given by (4) takes the form

$$\hat\mu = \sum_{j=1}^p \left(\frac{e^{\beta Z_j^2/\sigma^2}}{p^\alpha e^{2\beta}+e^{\beta Z_j^2/\sigma^2}}\,Z_j\right)v_j. \tag{9}$$

3. The performance

3.1. **A general risk bound.** The next result gives an upper bound on the L^2-risk of the estimation procedure. We remind the reader that $n\ge 3$ and set

$$\begin{array}{rccl}\phi: &]0,1[& \to &]0,+\infty[\\ & x & \mapsto & (x-1-\log x)/2\ ,\end{array} \tag{10}$$

which is decreasing.

Theorem 1. *Assume that β and N_* fulfill the condition*

$$\beta<1/4 \quad \textit{and} \quad N_*\ge 2+\frac{\log n}{\phi(4\beta)}, \tag{11}$$

with ϕ defined by (10). Assume also that $L_m\ge dim(\mathcal{S}_m)/2$, for all $m\in\mathcal{M}$. Then, we have the following upper bounds on the L^2-risk of the estimator $\hat\mu$

$$\begin{aligned}&\mathbb{E}\left(\|\mu-\hat\mu\|^2\right)\\ &\le\ -(1+\varepsilon_n)\,\frac{\bar\sigma^2}{\beta}\,\log\left[\sum_{m\in\mathcal{M}}\pi_m e^{-\beta\left[\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2-dim(\mathcal{S}_m)\sigma^2\right]/\bar\sigma^2-L_m}\right]+\frac{\sigma^2}{2\log n} &(12)\\ &\le\ (1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{\bar\sigma^2}{\beta}\left(L_m-\log\pi_m\right)\right\}+\frac{\sigma^2}{2\log n}, &(13)\end{aligned}$$

where $\varepsilon_n=(2n\log n)^{-1}$ and $\bar\sigma^2=\sigma^2+\|\mu-\Pi_{\mathcal{S}_*}\mu\|^2/N_*$.

The proof Theorem 1 is delayed to Section 6.3. Let us comment this result.

To start with, the Bound (12) may look somewhat cumberstone but it improves (13) when there are several good models to estimate μ. For example, we can derive from (12) the bound

$$\mathbb{E}\left(\|\mu-\hat{\mu}\|^2\right) \leq \\ (1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{\bar{\sigma}^2}{\beta}\left(L_m-\log\pi_m\right)\right\}+\inf_{\delta\geq 0}\left\{\delta-\frac{\bar{\sigma}^2}{\beta}\log|\mathcal{M}_\delta|\right\}+\frac{\sigma^2}{2\log n},$$

where $\mathcal{M}_\delta$ is the set made of those m^* in $\mathcal{M}$ fulfilling

$$\|\mu-\Pi_{\mathcal{S}_{m^*}}\mu\|^2+\frac{\bar{\sigma}^2}{\beta}\left(L_{m^*}-\log\pi_{m^*}\right)\leq\delta+\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{\bar{\sigma}^2}{\beta}\left(L_m-\log\pi_m\right)\right\}.$$

In the extreme case where all the quantities $\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{\bar{\sigma}^2}{\beta}\left(L_m-\log\pi_m\right)$ are equal, (12) then improves (13) by a factor $\beta^{-1}\bar{\sigma}^2\log|\mathcal{M}|$.

We now discuss the choice of the parameter β and the weights $\{L_m,\ m\in\mathcal{M}\}$. The choice $L_m=\dim(\mathcal{S}_m)/2$ seems to be the more accurate since it satisfies the conditions of Theorem 1 and minimizes the right hand side of (12) and (13). We shall mostly use this one in the following, but there are some cases where it is easier to use some (sharp) upper bound of the dimension of $\mathcal{S}_m$ instead of $\dim(\mathcal{S}_m)$ itself, see for example Section 4.3.

The largest parameter β fulfilling Condition (11) is

$$\beta=\frac{1}{4}\,\phi^{-1}\left(\frac{\log n}{N_*-2}\right)<\frac{1}{4}. \tag{14}$$

We suggest to use this value, since it minimizes the right hand side of (12) and (13). Nevertheless, as discussed in Section 3.2 for the situation of Section 2.2, it is sometimes possible to use larger values for β.

Finally, we would like to compare the bounds of Theorem 1 with the minimum over $\mathcal{M}$ of the risks given by (2). Roughly, the Bound (13) states that the estimator $\hat{\mu}$ achieves the best trade-off between the bias $\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2/\sigma^2$ and the complexity term $C_m=L_m-\log\pi_m$. More precisely, we derive from (13) the (cruder) bound

$$\mathbb{E}\left(\|\mu-\hat{\mu}\|^2\right)\leq(1+\varepsilon_n)\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2+\frac{1}{\beta}C_m\sigma^2\right\}+R_n^*\sigma^2, \tag{15}$$

with

$$\varepsilon_n=\frac{1}{2n\log n}\qquad\text{and}\qquad R_n^*=\frac{1}{2\log n}+\frac{\|\mu-\Pi_{\mathcal{S}_*}\mu\|^2}{\beta N^*\sigma^2}\sup_{m\in\mathcal{M}}C_m.$$

In particular, if C_m is of order $\dim(\mathcal{S}_m)$, then (15) allows to compare the risk of $\hat{\mu}$ with the infimum of the risks (2). We discuss this point in the following example.

Example. Assume that the family $\mathcal{M}$ has an index of complexity (M,a), as defined in [2], which means that

$$|\{m\in\mathcal{M},\ \dim(\mathcal{S}_m)=d\}|\leq Me^{ad},\quad\text{for all } d\geq 1.$$

If we choose, example given,

$$\pi_m = \frac{e^{-(a+1/2)\dim(\mathcal{S}_m)}}{\sum_{m'\in\mathcal{M}} e^{-(a+1/2)\dim(\mathcal{S}_{m'})}} \quad \text{and} \quad L_m = \dim(\mathcal{S}_m)/2, \tag{16}$$

we have $C_m \leq (a+1)\dim(\mathcal{S}_m) + \log(3M)$. Therefore, when β is given by (14) and $d_* \leq \kappa n$ for some $\kappa < 1$, we have

$$\mathbb{E}\left(\|\mu - \hat{\mu}\|^2\right) \leq (1+\varepsilon_n) \inf_{m\in\mathcal{M}} \left\{ \|\mu - \Pi_{\mathcal{S}_m}\mu\|^2 + \frac{a+1}{\beta}\dim(\mathcal{S}_m)\sigma^2 \right\} + R'_n\sigma^2,$$

with

$$R'_n = \frac{\log(3M)}{\beta} + \frac{1}{2\log n} + \frac{\|\mu - \Pi_{\mathcal{S}_*}\mu\|^2}{\sigma^2} \times \frac{(a+1)\kappa + n^{-1}\log(3M)}{\beta(1-\kappa)}\,.$$

In particular, for a given index of complexity (M, a) and a given κ, the previous bound gives an oracle inequality.

3.2. **On the choice of the parameter β.** The choice of the tuning parameter β is important in practice. Theorem 1 or Theorem 5 in [17] justify the choice of a β smaller than $1/4$. Nevertheless, Bayesian arguments [15] suggest to take a larger value for β, namely $\beta = 1/2$. In this section, we discuss this issue on the example of Section 2.2.

For the sake of simplicity, we will restrict to the case where the variance is known. We consider the weights (4) proposed by Leung and Barron, with the probability distribution π given by (7) with $\alpha = 1$, namely[1]

$$\pi_m = \left(1 + p^{-1}\right)^{-p} p^{-|m|}.$$

According to (9), the estimator $\hat{\mu}$ takes the form

$$\hat{\mu} = \sum_{j=1}^{p} s_\beta(Z_j/\sigma) Z_j\, v_j \quad \text{with } Z_j = < Y, v_j > \quad \text{and } s_\beta(z) = \frac{e^{\beta z^2}}{pe^{2\beta} + e^{\beta z^2}}. \tag{17}$$

To start with, we note that a choice $\beta > 1/2$ is not to be recommanded. Indeed, we can compare the shrinkage coefficient $s_\beta(Z_j/\sigma)$ to a threshold at level $T = (2 + \beta^{-1}\log p)\sigma^2$ since

$$s_\beta(Z_j/\sigma) \geq \frac{1}{2}\,\mathbf{1}_{\left\{Z_j^2 \geq T\right\}}.$$

For $\mu = 0$, the risk of $\hat{\mu}$ is then larger than a quarter of the risk of the threshold estimator $\hat{\mu}_T = \sum_{j=1}^{p} \mathbf{1}_{\left\{Z_j^2\geq T\right\}} Z_j\, v_j$, namely

$$\mathbb{E}\left(\|0 - \hat{\mu}\|^2\right) = \sum_{j=1}^{p} \mathbb{E}\left(s_\beta(Z_j/\sigma)^2 Z_j^2\right) \geq \frac{1}{4}\sum_{j=1}^{p} \mathbb{E}\left(\mathbf{1}_{\left\{Z_j^2\geq T\right\}} Z_j^2\right) = \frac{1}{4}\mathbb{E}\left(\|0 - \hat{\mu}_T\|^2\right).$$

[1]Note that this choice of α minimizes the rate of growth of $-\log\pi_m$ when p goes to infinity since

$$-\log\pi_m = p\log\left(1 + p^{-\alpha}\right) + \alpha|m|\log p = p^{1-\alpha} + \alpha|m|\log p + o\left(p^{1-\alpha}\right),$$

when π_m is given by (7) with $\alpha > 0$.

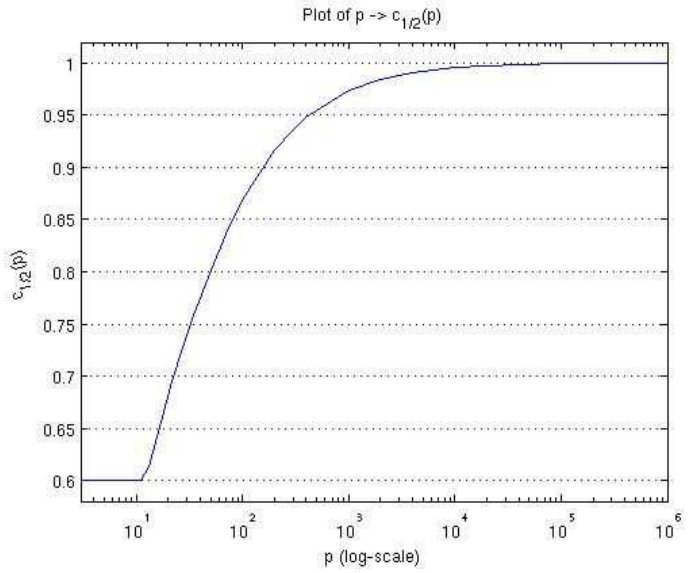

FIGURE 1. Plot of $p \mapsto c_{1/2}(p)$

Now, when the threshold T is of order $2K \log p$ with $K < 1$, the threshold estimator is known to behave poorly for $\mu = 0$, see [7] Section 7.2. Therefore, a choice $\beta > 1/2$ would give poor results at least when $\mu = 0$.

On the other hand, next proposition justifies the use of any $\beta \leq 1/2$ by a risk bound similar to (13). For $p \geq 1$ and $\beta > 0$, we introduce the numerical constants $\gamma_\beta(p) = \sqrt{2 + \beta^{-1} \log p}$ and

$$c_\beta(p) = \sup_{x \in [0, 4\gamma_\beta(p)]} \left[\frac{\int_{\mathbb{R}} \left(x - (x+z) s_\beta(x+z)\right)^2 e^{-z^2/2}\, dz/\sqrt{2\pi}}{\min(x^2, \gamma_\beta(p)^2) + \gamma_\beta(p)^2/p} \right] \vee 0.6.$$

This constant $c_\beta(p)$ can be numerically computed. For example, $c_{1/2}(p) \leq 1$ for any $3 \leq p \leq 10^6$, see Figure 1.

Proposition 2. *For $3 \leq p \leq n$ and $\beta \in [1/4, 1/2]$, the Euclidean risk of the estimator (17) is upper bounded by*

(18)
$$\mathbb{E}\left(\|\mu - \hat{\mu}\|^2\right) \leq \|\mu - \Pi_{\mathcal{S}_*}\mu\|^2 + c_\beta(p) \inf_{m \in \mathcal{M}} \left[\|\Pi_{\mathcal{S}_*}\mu - \Pi_{\mathcal{S}_m}\mu\|^2 + \left(2 + \beta^{-1} \log p\right)(|m| + 1)\sigma^2 \right].$$

The constant $c_\beta(p)$ is (crudely) bounded by 16 when $p \geq 3$ and $\beta \in [1/4, 1/2]$.

We delayed the proof of Proposition 2 to Section 6.4. We also emphasize that the bound $c_\beta(p) \leq 16$ is crude.

In light of the above result, $\beta = 1/2$ seems to be a good choice in this case and corresponds to the choice of Hartigan [15]. Note that the choice $\beta = 1/2$ has no reason to be a good choice in other situations. Indeed, a different choice of α in (7) would give different "good" values for β. For $\alpha \geq 1$, one may check that the "good" values for β are $\beta \leq \alpha/2$.

Remark 3. Proposition 2 does not provide an oracle inequality. The Bound (18) differs from the best trade off between the bias and the variance term by a $\log p$ factor. This is unavoidable from a minimax point of view as noticed in Donoho and Johnstone [13].

Remark 4. A similar analysis can be done for the estimator (8) when the variance is unknown. When the parameters α and b in (8) equal 1, we can justify the use of values of β fulfilling

$$\beta \leq \frac{1}{2}\ \phi^{-1}\left(\frac{\log p}{n-p}\right), \qquad \text{for } n > p \geq 3,$$

see the Appendix.

4. Choice of the models and the weights in different settings

In this section, we propose some choices of weights in three situations: the linear regression, the estimation of functions with bounded variation and regression in Besov spaces.

4.1. **Linear regression.**

We consider the case where the signal μ depends linearly on some observed explanatory variables $x^{(1)}, \ldots, x^{(p)}$, namely

$$\mu_i = \sum_{j=1}^{p} \theta_j x_i^{(j)}, \quad i = 1, \ldots, n.$$

The index i usually corresponds to an index of the experiment or to a time of observation. The number p of variables may be large, but we assume here that p is bounded by $n-3$.

4.1.1. *The case of ordered variables.* In some favorable situations, the explanatory variables $x^{(1)}, \ldots, x^{(p)}$ are naturally ordered. In this case, we will consider the models spanned by the m first explanatory variables, with m ranging from 0 to p. In this direction, we set $\mathcal{S}_0 = \{0\}$ and $\mathcal{S}_m = \text{span}\left\{x^{(1)}, \ldots, x^{(m)}\right\}$ for $m \in \{1, \ldots, p\}$, where $x^{(j)} = (x_1^{(j)}, \ldots, x_n^{(j)})'$. This collection of models is indexed by $\mathcal{M} = \{0, \ldots, p\}$ and contains one model per dimension. Note that $\mathcal{S}_*$ coincides here with $\mathcal{S}_p$.

We may use in this case the priors

$$\pi_m = \frac{e^{\alpha}-1}{e^{\alpha}-e^{-\alpha p}}\ e^{-\alpha m}, \quad m = 0, \ldots, p,$$

with $\alpha > 0$, set $L_m = m/2$ and takes the value (14) for β, with $N_* = n-p$. Then, according to Theorem 1 the performance of our procedure is controlled by

$$\begin{aligned}
&\mathbb{E}\left(\|\mu - \hat{\mu}\|^2\right) \\
&\quad \leq \quad (1+\varepsilon_n) \inf_{m \in \mathcal{M}} \left\{ \|\mu - \Pi_{\mathcal{S}_m}\mu\|^2 + \frac{\bar{\sigma}^2}{\beta}\,(\alpha + 1/2)m \right\} + 1.2\,\frac{\bar{\sigma}^2}{\beta}\,\log\left(\frac{e^{\alpha}}{e^{\alpha}-1}\right) + \frac{\sigma^2}{2\log n}.
\end{aligned}$$

with $\varepsilon_n = (2n\log n)^{-1}$ and $\bar\sigma^2 = \sigma^2 + \|\mu - \Pi_{\mathcal{S}_p}\mu\|^2/(n-p)$. As mentioned at the end of Section 3.1, the previous bound can be formulated as an oracle inequality when imposing the condition $p \le \kappa n$, for some $\kappa < 1$.

4.1.2. *The case of unordered variables.* When there is no natural order on the explanatory variables, we have to consider a larger collection of models. Set some $q \le p$ which represents the maximal number of explanatory variables we want to take into account. Then, we write $\mathcal{M}$ for all the subsets of $\{1,\dots,p\}$ of size less than q and $\mathcal{S}_m = \operatorname{span}\{x^{(j)},\ j\in m\}$ for any nonempty m. We also set $\mathcal{S}_\emptyset = \{0\}$ and $\mathcal{S}_* = \operatorname{span}\{x^{(1)},\dots,x^{(p)}\}$. Note that the cardinality of $\mathcal{M}$ is of order p^q, so when p is large the value q should remain small in practice.

A possible choice for π_m is

$$\pi_m = \left[\binom{p}{|m|}(|m|+1)H_q\right]^{-1} \quad \text{with} \quad H_q = \sum_{d=0}^{q}\frac{1}{d+1} \ \le\ 1+\log(q+1).$$

Again, we choose the value (14) for β with $N_* = n-p$ and $L_m = |m|/2$. With this choice, combining the inequality $\binom{p}{|m|} \le (e|m|/p)^{|m|}$ with Theorem 1 gives the following bound on the risk of the procedure

$$\begin{aligned}
&\mathbb{E}\left(\|\mu-\hat\mu\|^2\right)\\
&\le\ [1+\varepsilon_n]\inf_{m\in\mathcal{M}}\left\{\|\mu-\Pi_{\mathcal{S}_m}\mu\|^2 + \frac{\bar\sigma^2}{\beta}\left[|m|\left(3/2+\log\frac{p}{|m|}\right)+\log(|m|+1)\right]\right\}\\
&\quad +\frac{\sigma^2}{2\log n} + \frac{\bar\sigma^2}{\beta}\log\log[(q+1)e]\,,
\end{aligned}$$

with $\varepsilon_n = (2n\log n)^{-1}$ and $\bar\sigma^2 = \sigma^2 + \|\mu-\Pi_{\mathcal{S}_*}\mu\|^2/(n-p)$.

Remark. When the family $\{x^{(1)},\dots,x^{(p)}\}$ is orthogonal and $q=p$, we fall into the setting of Section 2.2. An alternative in this case is to use $\hat\mu$ given by (8), which is easy to compute numerically.

4.2. **Estimation of BV functions.** We consider here the functional setting

$$\mu_i = f(x_i), \quad i=1,\dots,n \tag{19}$$

where $f:[0,1]\to\mathbb{R}$ is an unknown function and $x_1,\dots,x_n$ are n deterministic points of $[0,1]$. We assume for simplicity that $0 = x_1 < x_2 < \cdots < x_n < x_{n+1} = 1$ and $n = 2^{J_n} \ge 8$. We set $J^* = J_n - 1$ and $\Lambda^* = \cup_{j=0}^{J^*}\Lambda(j)$ with $\Lambda(0) = \{(0,0)\}$ and $\Lambda(j) = \{j\}\times\{0,\dots,2^{j-1}-1\}$ for $j\ge 1$. For $(j,k)\in\Lambda^*$ we define $v_{j,k}\in\mathbb{R}^n$ by

$$[v_{j,k}]_i = 2^{(j-1)/2}\left(\mathbf{1}_{I^+_{j,k}}(i) - \mathbf{1}_{I^-_{j,k}}(i)\right),\quad i=1,\dots,n$$

with $I^+_{j,k} = \{1+(2k+1)2^{-j}n,\dots,(2k+2)2^{-j}n\}$ and $I^-_{j,k} = \{1+2k2^{-j}n,\dots,(2k+1)2^{-j}n\}$. The family $\{v_{j,k},\ (j,k)\in\Lambda^*\}$ corresponds to the image of the points $x_1,\dots,x_n$ by a Haar basis

(see Section 6.5) and it is orthonormal for the scalar product

$$< x, y >_n = \frac{1}{n} \sum_{i=1}^{n} x_i y_i.$$

We use the collection of models $\mathcal{S}_m = \operatorname{span}\{v_{j,k},\ (j,k) \in m\}$ indexed by $\mathcal{M} = \mathcal{P}(\Lambda^*)$ and fall into the setting of Section 2.2. We choose the distribution π given by (7) with $p = n/2$ and $\alpha = 1$. We also set $b = 1$ and take some β fulfilling

$$\beta \leq \frac{1}{2}\phi^{-1}\left(\frac{2\log(n/2)}{n}\right).$$

According to Proposition 1 the estimator (6) then takes the form

$$\hat{\mu} = \sum_{j=0}^{J^*} \sum_{k \in \Lambda(j)} \left(\frac{Z_{j,k} \exp\left(n\beta Z_{j,k}^2/\hat{\sigma}^2\right)}{en/2 + \exp\left(n\beta Z_{j,k}^2/\hat{\sigma}^2\right)} \right) v_{j,k}\ , \tag{20}$$

with $Z_{j,k} =< Y, v_{j,k} >_n$ and

$$\hat{\sigma}^2 = 2\left(< Y, Y >_n^2 - \sum_{j=0}^{J^*} \sum_{k \in \Lambda(j)} Z_{j,k}^2 \right).$$

Next corollary gives the rate of convergence of this estimator when f has bounded variation, in terms of the norm $\|\cdot\|_n$ induced by the scalar product $< \cdot, \cdot >_n$.

Corollary 1. *In the setting described above, there exists a numerical constant C such that for any function f with bounded variation $V(f)$*

$$\mathbb{E}\left(\|\mu - \hat{\mu}\|_n^2\right) \leq C \max\left\{ \left(\frac{V(f)\sigma^2 \log n}{n}\right)^{2/3}\ ,\ \frac{V(f)^2}{n}\ ,\ \frac{\sigma^2 \log n}{n} \right\}.$$

The proof is delayed to Section 6.5. The minimax rate in this setting is $(V(f)\sigma^2/n)^{2/3}$. So, the rate of convergence of the estimator differs from the minimax rate by a $(\log n)^{2/3}$ factor. We can actually obtain a rate-minimax estimator by using a smaller collection of models similar to the one introduced in the next section, but we lose then Formula (20).

4.3. **Regression on Besov space $\mathcal{B}^{\alpha}_{p,\infty}[0,1]$.** We consider again the setting (19) with $f : [0,1] \to \mathbb{R}$ and introduce a $L^2([0,1], dx)$-orthonormal family $\left\{\phi_{j,k}, j \geq 0, k = 1 \dots 2^j\right\}$ of compactly support wavelets with regularity r. We will use models generated by finite subsets of wavelets. If we want that our estimator shares some good adaptive properties on Besov spaces, we shall introduce a family of models induced by the compression algorithm of Birgé and Massart [7]. This collection turns to be slightly more intricate than the family used in the previous section. We start with some $\kappa < 1$ and set $J_* = \lfloor \log(\kappa n/2)/\log 2 \rfloor$. The largest approximation space we will consider is

$$\mathcal{F}_* = \operatorname{span}\left\{\phi_{j,k}, j = 0 \dots J_*, k = 1 \dots 2^j\right\},$$

whose dimension is bounded by κn. For $1 \le J \le J_*$, we define

$$\mathcal{M}_J = \left\{ m = \bigcup_{j=0}^{J_*} \{j\} \times A_j, \text{ with } A_j \in \Lambda_{j,J} \right\},$$

where $\Lambda_{j,J} = \big\{\big\{1, \dots, 2^j\big\}\big\}$ when $j \le J-1$ and

$$\Lambda_{j,J} = \big\{A \subset \big\{1, \dots, 2^j\big\} : \ |A| = \lfloor 2^J/(j-J+1)^3 \rfloor \big\}, \quad \text{when } J \le j \le J_*.$$

To m in $\mathcal{M} = \cup_{J=1}^{J_*} \mathcal{M}_J$, we associate $\mathcal{F}_m = \operatorname{span}\{\phi_{j,k},\ (j,k) \in m\}$ and define the model $\mathcal{S}_m$ by

$$\mathcal{S}_m = \{(f(x_1), \dots, f(x_n)),\ f \in \mathcal{F}_m\} \subset \mathcal{S}_* = \{(f(x_1), \dots, f(x_n)),\ f \in \mathcal{F}_*\}.$$

When $m \in \mathcal{M}_J$, the dimension of $\mathcal{S}_m$ is bounded from above by

$$\dim(\mathcal{S}_m) \le \sum_{j=0}^{J-1} 2^j + \sum_{j=J}^{J_*} \frac{2^J}{(j-J+1)^3} \le 2^J \left[1 + \sum_{k=1}^{J_*-J+1} k^{-3} \right] \le 2.2 \cdot 2^J \tag{21}$$

and $\dim(\mathcal{S}_*) \le \kappa n$. Note also that the cardinality of $\mathcal{M}_J$ is

$$|M_J| = \prod_{j=J}^{J_*} \binom{2^j}{\lfloor 2^J/(j-J+1)^3 \rfloor}.$$

To estimate μ, we use the estimator $\hat{\mu}$ given by (6) with β given by (14) and

$$L_m = 1.1 \cdot 2^J \quad \text{and} \quad \pi_m = \left[2^J (1 - 2^{J_*}) \prod_{j=J}^{J_*} \binom{2^j}{\lfloor 2^J/(j-J+1)^3 \rfloor} \right]^{-1}, \quad \text{for } m \in \mathcal{M}_J.$$

Next corollary gives the rate of convergence of the estimator $\hat{\mu}$ when f belongs to some Besov ball $\mathcal{B}^\alpha_{p,\infty}(R)$ with $1/p < \alpha < r$ (we refer to De Vore and Lorentz [12] for a precise definition of Besov spaces). As it is usual in this setting, we express the result in terms of the norm $\|\cdot\|_n^2 = \|\cdot\|^2/n$ on $\mathbb{R}^n$.

Corollary 2. *For any $p, R > 0$ and $\alpha \in]1/p, r[$, there exists some constant C not depending on n and σ^2 such that the estimator $\hat{\mu}$ defined above fulfills*

$$\mathbb{E}\left(\|\mu - \hat{\mu}\|_n^2\right) \le C \max\left\{ \left(\frac{\sigma^2}{n}\right)^{2\alpha/(2\alpha+1)}, \ \frac{1}{n^{2(\alpha-1/p)}}, \ \frac{\sigma^2}{n} \right\}$$

for any μ given by (19) with $f \in \mathcal{B}^\alpha_{p,\infty}(R)$.

The proof is delayed to Section 6.6. We remind that the rate $\left(\sigma^2/n\right)^{2\alpha/(2\alpha+1)}$ is minimax in this framework, see Yang and Barron [24].

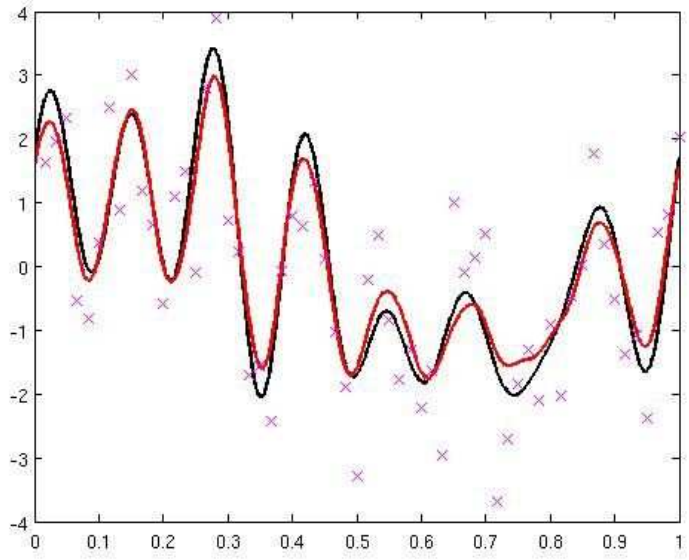

FIGURE 2. Recovering a signal from noisy observations (the crosses). The signal is in black, the estimator is in red.

5. A numerical illustration

We illustrate the use of our procedure on a numerical simulation. We start from a signal

$$f(x) = 0.7\cos(x) + \cos(7x) + 1.5\sin(x) + 0.8\sin(5x) + 0.9\sin(8x), \quad x \in [0,1]$$

which is in black in Figure 2. We have $n = 60$ noisy observations of this signal

$$Y_i = f(x_i) + \sigma\varepsilon_i, \quad i = 1, \dots, 60, \qquad \text{(the crosses in Figure 2)}$$

where $x_i = i/60$ and $\varepsilon_1, \dots, \varepsilon_{60}$ are 60 i.i.d. standard Gaussian random variables. The noise level σ is not known (σ equals 1 in Figure 2). To estimate f we will expand the observations on the Fourier basis

$$\{1, \cos(2\pi x), \dots, \cos(40\pi x), \sin(2\pi x), \dots, \sin(40\pi x)\}.$$

In this direction, we introduce the $p = 41$ vectors $v_1, \dots, v_{41}$ given by

$$v_j = \begin{cases} \left(\sqrt{\frac{2}{n}}\sin(2\pi j x_1), \dots, \sqrt{\frac{2}{n}}\sin(2\pi j x_n)\right)' & \text{when } j \in \{1, \dots, 20\} \\ \left(\sqrt{\frac{1}{n}}, \dots, \sqrt{\frac{1}{n}}\right)' & \text{when } j = 21 \\ \left(\sqrt{\frac{2}{n}}\cos(2\pi(j-21)x_1), \dots, \sqrt{\frac{2}{n}}\cos(2\pi(j-21)x_n)\right)' & \text{when } j \in \{22, \dots, 41\}. \end{cases}$$

This vectors $\{v_1, \dots, v_{41}\}$ form an orthonormal family for the usual scalar product of $\mathbb{R}^n$ and we fall into the setting of Section 2.2.

We estimate $(f(x_1),\dots,f(x_n))'$ with $\hat\mu$ given by (8) with the parameter $\alpha = 1$, $b = 1$ and $\beta = 1/3$. Finally, we estimate f with

$$\hat f(x) = \hat a_0 + \sum_{j=1}^{20} \hat a_j \cos(2\pi j x) + \sum_{j=1}^{20} \hat b_j \sin(2\pi j x) \qquad \text{(in red in Figure 2)}$$

where $\hat a_0 = \sqrt{(1/n)} < \hat\mu, v_{21} >$ and $\hat a_j = \sqrt{(2/n)} < \hat\mu, v_{j+21} >$, $\hat b_j = \sqrt{(2/n)} < \hat\mu, v_j >$ for $j = 1,\dots,20$. The plot of $\hat f$ is in red in Figure 2.

6. Proofs

6.1. **Proof of Proposition 1.** Let us first express the weights w_m in terms of the Z_js and $\tau = \alpha\log p + b$. We use below the convention, that the sum of zero term is equal to zero. Then for any $m \in \mathcal{M}$, we have

$$\begin{aligned} w_m &= \frac{\exp\left(\beta\|\hat\mu_m\|^2/\hat\sigma^2 - (\alpha\log p + b)|m|\right)}{\sum_{m'\in\mathcal{M}} \exp\left(\beta\|\hat\mu_{m'}\|^2/\hat\sigma^2 - (\alpha\log p + b)|m'|\right)} \\ &= \frac{\exp\left(\sum_{k\in m}\left(\beta Z_k^2/\hat\sigma^2 - \tau\right)\right)}{\sum_{m'\in\mathcal{M}} \exp\left(\sum_{k\in m'}\left(\beta Z_k^2/\hat\sigma^2 - \tau\right)\right)}\,. \end{aligned}$$

We write $\mathcal{M}_j$ for the set of all the subsets of $\{1,\dots,j-1,j+1,\dots,p\}$. Then, for any $j \in \{1,\dots,p\}$

$$\begin{aligned} c_j &= \sum_{m\in\mathcal{M}} \mathbb{1}_{j\in m} w_m \\ &= \frac{\sum_{m\in\mathcal{M}} \mathbb{1}_{j\in m} \exp\left(\sum_{k\in m}\left(\beta Z_k^2/\hat\sigma^2 - \tau\right)\right)}{\sum_{m\in\mathcal{M}} \exp\left(\sum_{k\in m}\left(\beta Z_k^2/\hat\sigma^2 - \tau\right)\right)}. \end{aligned}$$

Note that any subset $m \in \mathcal{M}$ with j inside can be written as $\{j\}\cup m'$ with $m' \in \mathcal{M}_j$, so that

$$\begin{aligned} c_j &= \frac{e^{\beta Z_j^2/\hat\sigma^2 - \tau}\sum_{m'\in\mathcal{M}_j} \exp\left(\sum_{k\in m'}\left(\beta Z_k^2/\hat\sigma^2 - \tau\right)\right)}{e^{\beta Z_j^2/\hat\sigma^2 - \tau}\sum_{m'\in\mathcal{M}_j} \exp\left(\sum_{k\in m'}\left(\beta Z_k^2/\hat\sigma^2 - \tau\right)\right) + \sum_{m\in\mathcal{M}_j} \exp\left(\sum_{k\in m}\left(\beta Z_k^2/\hat\sigma^2 - \tau\right)\right)} \\ &= \frac{e^{\beta Z_j^2/\hat\sigma^2 - \tau}}{e^{\beta Z_j^2/\hat\sigma^2 - \tau} + 1}. \end{aligned}$$

Formula (8) follows.

6.2. **A preliminary lemma.**

Lemma 1. *Consider an integer N larger than 2 and a random variable X, such that NX is distributed as a χ^2 of dimension N. Then, for any $0 < a < 1$,*

$$\mathbb{E}\left[(a - X)_+\right] \le \mathbb{E}\left[\left(\frac{a}{X} - 1\right)_+\right] \le \frac{2}{(1-a)(N-2)}\exp\left(-N\phi(a)\right) \tag{22}$$

with $\phi(a) = \frac{1}{2}(a - 1 - \log a) > \frac{1}{4}(1-a)^2$.

Proof: Remind first that when $g : \mathbb{R}^+ \to \mathbb{R}^+$ and $f : \mathbb{R}^+ \to \mathbb{R}^+$ have opposite monotonicity,

$$\mathbb{E}[g(X)f(X)] \leq \mathbb{E}[g(X)]\,\mathbb{E}[f(X)].$$

Setting $g(x) = x$ and $f(x) = (a/x - 1)_+$ leads to the first inequality since $\mathbb{E}(X) = 1$.

We now turn to the second inequality. To start with, we note that

$$\begin{aligned}\mathbb{E}\left[\left(\frac{a}{X} - 1\right)_+\right] &= a\mathbb{E}\left[\left(\frac{1}{X} - \frac{1}{a}\right)_+\right] \\ &= a\int_{1/a}^{+\infty} \mathbb{P}\left(X \leq \frac{1}{t}\right)\, dt\,.\end{aligned}$$

Markov inequality gives for any $\lambda \geq 0$

$$\mathbb{P}\left(X \leq \frac{1}{t}\right) \leq e^{\lambda/t}\mathbb{E}\left(e^{-\lambda X}\right) = e^{\lambda/t}\left(1 + \frac{2\lambda}{N}\right)^{-N/2}$$

and choosing $\lambda = N(t-1)/2$ leads to

$$\mathbb{P}\left(X \leq \frac{1}{t}\right) \leq \frac{1}{t^{N/2}}\,\exp\left(\frac{N}{2}(1 - 1/t)\right) \leq \exp\left(-N\phi(1/t)\right) \tag{23}$$

for any $t > 1$. Putting pieces together ensures the bound

$$\begin{aligned}\mathbb{E}\left[\left(\frac{a}{X} - 1\right)_+\right] &\leq a\int_{1/a}^{+\infty} \exp\left(\frac{N}{2}(1 - 1/t)\right)\frac{dt}{t^{N/2}} \\ &\leq a\int_0^a \exp\left(\frac{N}{2}(1-x)\right)x^{N/2-2}dx\end{aligned}$$

for any $0 < a < 1$. Iterating integrations by parts leads to

$$\begin{aligned}\mathbb{E}\left[\left(\frac{a}{X} - 1\right)_+\right] &= \exp\left(\frac{N}{2}(1-a)\right)\frac{2\,a^{N/2}}{N-2}\sum_{k\geq 0}\frac{a^k(N/2)^k}{\prod_{i=0}^{k-1}(N/2+i)} \\ &\leq \exp\left(\frac{N}{2}(1-a)\right)\frac{2\,a^{N/2}}{(N-2)(1-a)} \\ &\leq \frac{2}{(1-a)(N-2)}\exp\left(-N\phi(a)\right)\end{aligned}$$

for $0 < a < 1$, and the bound (22) follows.

6.3. **Proof of Theorem 1.** To keep formulas short, we write d_m for the dimension of $\mathcal{S}_m$, and use the following notations for the various projections

$$\hat{\mu}_* = \Pi_{\mathcal{S}_*}Y, \quad \mu_* = \Pi_{\mathcal{S}_*}\mu \quad \text{and} \quad \mu_m = \Pi_{\mathcal{S}_m}\mu, \quad m \in \mathcal{M}.$$

It is also convenient to write w_m in the form

$$w_m = \frac{\pi_m}{\mathcal{Z}'}\exp\left(-\beta\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\hat{\sigma}^2} - L_m\right)$$

with $\mathcal{Z}' = e^{-\beta\|\hat{\mu}_*\|^2/\hat{\sigma}^2}\mathcal{Z}$.

By construction, the estimator $\hat{\mu}$ belongs to $\mathcal{S}_*$, so according to Pythagorean equality we have $\|\mu-\hat{\mu}\|^2 = \|\mu-\mu_*\|^2 + \|\mu_*-\hat{\mu}\|^2$. The first part is non random, and we only need to control the second part.

According to Theorem 1 in Leung and Barron [17], the Stein's unbiased estimate of the L^2-risk $\mathbb{E}\left(\|\mu_*-\hat{\mu}\|^2\right)/\sigma^2$ of the estimator $\hat{\mu}$ on $\mathcal{S}_*$ can be written as

$$S\left(\hat{\mu}\right) = \sum_{m\in\mathcal{M}} w_m \left[\frac{\|\hat{\mu}_*-\hat{\mu}_m\|^2}{\sigma^2} + 2d_m - d_* - \frac{\|\hat{\mu}_m-\hat{\mu}\|^2}{\sigma^2} + 2\beta\nabla\left(\frac{\|\hat{\mu}_*-\hat{\mu}_m\|^2}{\hat{\sigma}^2}\right).\left(\hat{\mu}-\hat{\mu}_m\right)\right].$$

When expanding the gradient into

$$\nabla\left(\frac{\|\hat{\mu}_*-\hat{\mu}_m\|^2}{\hat{\sigma}^2}\right).\left(\hat{\mu}-\hat{\mu}_m\right) = \frac{2\left(\hat{\mu}_*-\hat{\mu}_m\right)}{\hat{\sigma}^2}.\left(\hat{\mu}-\hat{\mu}_m\right) + \|\hat{\mu}_*-\hat{\mu}_m\|^2\nabla\left(1/\hat{\sigma}^2\right).\left(\hat{\mu}-\hat{\mu}_m\right),$$

the term $\|\hat{\mu}_*-\hat{\mu}_m\|^2\nabla\left(1/\hat{\sigma}^2\right).\left(\hat{\mu}-\hat{\mu}_m\right)$ turns to be 0, since $\nabla\left(1/\hat{\sigma}^2\right)$ is orthogonal to $\mathcal{S}_*$. Furthermore, the sum

$$\sum_{m\in\mathcal{M}} w_m\left(\hat{\mu}-\hat{\mu}_*\right).\left(\hat{\mu}-\hat{\mu}_m\right)$$

also equals 0, so an unbiased estimate of $\mathbb{E}\left(\|\mu_*-\hat{\mu}\|^2\right)/\sigma^2$ is

$$S\left(\hat{\mu}\right) = \sum_{m\in\mathcal{M}} w_m\left[\frac{\|\hat{\mu}_*-\hat{\mu}_m\|^2}{\sigma^2} + 2d_m + \left(4\beta-\frac{\hat{\sigma}^2}{\sigma^2}\right)\frac{\|\hat{\mu}_m-\hat{\mu}\|^2}{\hat{\sigma}^2}\right] - d_*.$$

We control the last term thanks to the upper bound

$$\begin{aligned}\sum_{m\in\mathcal{M}} w_m\|\hat{\mu}_m-\hat{\mu}\|^2 &= \sum_{m\in\mathcal{M}} w_m\|\hat{\mu}_*-\hat{\mu}_m\|^2 - \|\hat{\mu}-\hat{\mu}_*\|^2\\ &\le \sum_{m\in\mathcal{M}} w_m\|\hat{\mu}_*-\hat{\mu}_m\|^2\end{aligned}$$

and get

$$\begin{aligned}S\left(\hat{\mu}\right) &\le \left[\frac{\hat{\sigma}^2}{\sigma^2} + \left(4\beta-\frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right]\sum_{m\in\mathcal{M}} w_m\,\frac{\|\hat{\mu}_*-\hat{\mu}_m\|^2}{\hat{\sigma}^2} + 2\sum_{m\in\mathcal{M}} w_m d_m - d_*\\ &\le \left[\frac{\hat{\sigma}^2}{\sigma^2} + \left(4\beta-\frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right]\sum_{m\in\mathcal{M}} w_m\left[\frac{\|\hat{\mu}_*-\hat{\mu}_m\|^2}{\hat{\sigma}^2} + \frac{L_m}{\beta}\right] - d_*\\ &\quad + \sum_{m\in\mathcal{M}} w_m\left(2d_m - \left[\frac{\hat{\sigma}^2}{\sigma^2} + \left(4\beta-\frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right]\frac{L_m}{\beta}\right)\end{aligned}$$

where $(x)_+ = \max(0,x)$. First note that when $L_m \ge d_m/2$ we have

$$\begin{aligned}2d_m - \left[\frac{\hat{\sigma}^2}{\sigma^2} + \left(4\beta-\frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right]\frac{L_m}{\beta} &\le \left[2 - \frac{\hat{\sigma}^2}{2\beta\sigma^2} - \frac{1}{2\beta}\left(4\beta-\frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right]d_m\\ &\le \min\left(0, 2-\frac{\hat{\sigma}^2}{2\beta\sigma^2}\right)d_m \;\le\; 0.\end{aligned}$$

Therefore, setting $\hat{\delta}_\beta = \left(4\beta\sigma^2/\hat{\sigma}^2 - 1\right)_+$ we get

$$S(\hat{\mu}) \leq \frac{\hat{\sigma}^2}{\sigma^2}\left(1+\hat{\delta}_\beta\right)\sum_{m\in\mathcal{M}} w_m\left[\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\hat{\sigma}^2} + \frac{L_m}{\beta}\right] - d_*.$$

Let us introduce the Kullback divergence between two probability distributions $\{\alpha_m,\ m\in\mathcal{M}\}$ and $\{\pi_m,\ m\in\mathcal{M}\}$ on $\mathcal{M}$

$$\mathcal{D}(\alpha|\pi) = \sum_{m\in\mathcal{M}} \alpha_m \log\frac{\alpha_m}{\pi_m} \ \geq\ 0$$

and the function

$$\mathcal{E}^\pi_\beta(\alpha) = \sum_{m\in\mathcal{M}} \alpha_m\left[\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\hat{\sigma}^2} + \frac{L_m}{\beta}\right] + \frac{1}{\beta}\mathcal{D}(\alpha|\pi).$$

The latter function is convex on the simplex $S^+_{\mathcal{M}} = \left\{\alpha\in[0,1]^{|\mathcal{M}|},\ \sum_{m\in\mathcal{M}}\alpha_m = 1\right\}$ and can be interpreted as a free energy function. Therefore, it is minimal for the Gibbs measure $\{w_m,\ m\in\mathcal{M}\}$ and for any $\alpha\in S^+_{\mathcal{M}}$,

$$\begin{aligned} S(\hat{\mu}) &\leq \frac{\hat{\sigma}^2}{\sigma^2}\left(1+\hat{\delta}_\beta\right)\left[\sum_{m\in\mathcal{M}}\alpha_m\left[\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\hat{\sigma}^2} + \frac{L_m}{\beta}\right] + \frac{1}{\beta}\mathcal{D}(\alpha|\pi) - \frac{1}{\beta}\mathcal{D}(w|\pi)\right] - d_* \\ &\leq \left(1+\hat{\delta}_\beta\right)\left[\sum_{m\in\mathcal{M}}\alpha_m\left[\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\sigma^2} + \frac{\hat{\sigma}^2}{\beta\sigma^2}L_m\right] + \frac{\hat{\sigma}^2}{\beta\sigma^2}\mathcal{D}(\alpha|\pi)\right] - d_*. \end{aligned}$$

We fix a probability distribution $\alpha\in S^+_{\mathcal{M}}$ and take the expectation in the last inequality to get

$$\begin{aligned} &\mathbb{E}\left[S(\hat{\mu})\right] \leq \\ &\quad\left(1+\mathbb{E}(\hat{\delta}_\beta)\right)\sum_{m\in\mathcal{M}}\alpha_m\mathbb{E}\left[\frac{\|\hat{\mu}_* - \hat{\mu}_m\|^2}{\sigma^2}\right] + \mathbb{E}\left[\frac{\hat{\sigma}^2}{\beta\sigma^2}(1+\hat{\delta}_\beta)\right]\left[\sum_{m\in\mathcal{M}}\alpha_m L_m + \mathcal{D}(\alpha|\pi)\right] - d_*. \end{aligned}$$

Since $\hat{\sigma}^2/\sigma^2$ is stochastically larger than a random variable X with $\chi^2(N_*)/N_*$ distribution, Lemma 1 ensures that the two expectations

$$\mathbb{E}\left[\frac{\hat{\sigma}^2}{\sigma^2}\hat{\delta}_\beta\right] = \mathbb{E}\left[\left(4\beta - \frac{\hat{\sigma}^2}{\sigma^2}\right)_+\right] \quad\text{and}\quad \mathbb{E}\left[\hat{\delta}_\beta\right] = \mathbb{E}\left[\left(\frac{4\beta\sigma^2}{\hat{\sigma}^2} - 1\right)_+\right]$$

are bounded by

$$\frac{2}{(1-4\beta)(N_*-2)}\exp\left(-N_*\phi(4\beta)\right),$$

with $\phi(x) = (x - 1 - \log(x))/2$. Furthermore, the condition $N_* \geq 2 + (\log n)/\phi(4\beta)$ enforces

$$\frac{2}{(1-4\beta)(N_*-2)}\exp\left(-N_*\phi(4\beta)\right) \ \leq\ \frac{2\phi(4\beta)e^{-2\phi(4\beta)}}{(1-4\beta)n\log n} \ \leq\ \frac{1}{2n\log n} \ =\ \varepsilon_n.$$

Putting pieces together, we obtain

$$\begin{aligned}
&\frac{\mathbb{E}\left[\|\mu-\hat\mu\|^2\right]}{\sigma^2}\\
&= \frac{\|\mu-\mu_*\|^2}{\sigma^2} + \mathbb{E}[S(\hat\mu)]\\
&\le \frac{\|\mu-\mu_*\|^2}{\sigma^2} + (1+\varepsilon_n)\sum_{m\in\mathcal{M}}\alpha_m\mathbb{E}\left[\frac{\|\hat\mu_*-\hat\mu_m\|^2}{\sigma^2}\right]\\
&\quad + \left(\frac{\bar\sigma^2}{\beta\sigma^2}+\varepsilon_n\right)\left[\sum_{m\in\mathcal{M}}\alpha_m L_m + \mathcal{D}(\alpha|\pi)\right] - d_*\\
&\le \frac{\|\mu-\mu_*\|^2}{\sigma^2} + (1+\varepsilon_n)\sum_{m\in\mathcal{M}}\alpha_m\left[\frac{\|\mu_*-\mu_m\|^2}{\sigma^2} + d_* - d_m + \frac{\bar\sigma^2}{\beta\sigma^2}(L_m+\mathcal{D}(\alpha|\pi))\right] - d_*\\
&\le (1+\varepsilon_n)\left[\sum_{m\in\mathcal{M}}\alpha_m\left[\frac{\|\mu-\mu_m\|^2}{\sigma^2} - d_m + \frac{\bar\sigma^2}{\beta\sigma^2}L_m\right] + \frac{\bar\sigma^2}{\beta\sigma^2}\mathcal{D}(\alpha|\pi)\right] + \varepsilon_n d_*.
\end{aligned}$$

This inequality holds for any non-random probability distribution $\alpha\in S^+_{\mathcal{M}}$, so it holds in particular for the Gibbs measure

$$\alpha_m = \frac{\pi_m}{\mathcal{Z}_\beta}\exp\left[-\frac{\beta}{\bar\sigma^2}\left(\|\mu-\mu_m\|^2 - d_m\sigma^2\right) - L_m\right],\quad m\in\mathcal{M}$$

where $\mathcal{Z}_\beta$ normalizes the sum of the α_ms to one. For this choice of α_m we obtain

$$\frac{\mathbb{E}\left[\|\mu-\hat\mu\|^2\right]}{\sigma^2} \le -\frac{(1+\varepsilon_n)\bar\sigma^2}{\beta\sigma^2}\log\left[\sum_{m\in\mathcal{M}}\pi_m\exp\left[-\frac{\beta}{\bar\sigma^2}\left(\|\mu-\mu_m\|^2 - d_m\sigma^2\right) - L_m\right]\right] + \varepsilon_n d_*$$

which ensures (12) since $d_*\le n$. To get (13) simply note that

$$\begin{aligned}
&\sum_{m\in\mathcal{M}}\pi_m\exp\left[-\frac{\beta}{\bar\sigma^2}\left(\|\mu-\mu_m\|^2 - d_m\sigma^2\right) - L_m\right]\\
&\qquad\qquad \ge \exp\left[-\frac{\beta}{\bar\sigma^2}\left(\|\mu-\mu_{m^*}\|^2 - d_{m^*}\sigma^2\right) - L_{m^*} - \log\pi_{m^*}\right]
\end{aligned}$$

for any $m^*\in\mathcal{M}$.

6.4. **Proof of Proposition 2.** We use along the proof the notations $\mu_* = \Pi_{\mathcal{S}_*}\mu$, $[x]_+ = \max(x,0)$ and $\gamma = \gamma_\beta(p) = \sqrt{2+\beta^{-1}\log p}$. We omit the proof of the bound $c_\beta(p)\le 16$ for $\beta\in[1/4,1/2]$ and $p\ge 3$. This proof (of minor interest) can be found in the Appendix.

The risk of $\hat\mu$ is given by

$$\mathbb{E}\left(\|\mu-\hat\mu\|^2\right) = \|\mu-\mu_*\|^2 + \sum_{j=1}^p \mathbb{E}\left(\left(<\mu,v_j> - s_\beta(Z_j/\sigma)Z_j\right)^2\right).$$

Note that $Z_j =< \mu, v_j > +\sigma < \varepsilon, v_j >$, with $< \varepsilon, v_j >$ distributed as a standard Gaussian random variable. As a consequence, when $| < \mu, v_j > | \leq 4\gamma\sigma$ we have

$$\mathbb{E}\left((< \mu, v_j > -s_\beta(Z_j/\sigma)Z_j)^2 \right) \leq c_\beta(p) \left[\min(< \mu, v_j >^2, \gamma^2\sigma^2) + \gamma^2\sigma^2/p \right]. \tag{24}$$

If we prove the same inequality for $| < \mu, v_j > | \geq 4\gamma\sigma$, then

$$\begin{aligned} \mathbb{E}\left(\|\mu - \hat{\mu}\|^2 \right) &\leq \|\mu - \mu_*\|^2 + c_\beta(p) \sum_{j=1}^{p} \min(< \mu, v_j >^2, \gamma^2\sigma^2) + \gamma^2\sigma^2 \\ &\leq \|\mu - \mu_*\|^2 + c_\beta(p) \inf_{m \in \mathcal{M}} \left[\|\mu_* - \mu_m\|^2 + \gamma^2(|m| + 1)\sigma^2 \right]. \end{aligned}$$

This last inequality is exactly the Bound (18).

To conclude the proof of Proposition 2, we need to check that

$$\mathbb{E}\left((x - s_\beta(x + Z)(x + Z))^2 \right) \leq c_\beta(p)\gamma^2 \qquad \text{for } x \geq 4\gamma,$$

where Z is distributed as a standard Gaussian random variable. We first note that

$$\begin{aligned} \mathbb{E}\left((x - s_\beta(x + Z)(x + Z))^2 \right) &\leq 2x^2\mathbb{E}\left((1 - s_\beta(x + Z))^2 \right) + 2\mathbb{E}(s_\beta(x + Z)^2 Z^2) \\ &\leq 2x^2\mathbb{E}\left((1 - s_\beta(x + Z))^2 \right) + 2. \end{aligned}$$

We can bound $(1 - s_\beta(x + Z))^2$ as follows

$$\begin{aligned} (1 - s_\beta(x + Z))^2 &= \left(\frac{1}{1 + \exp\left(\beta\left[(x + Z)^2 - \gamma^2\right]\right)} \right)^2 \\ &\leq \exp\left(-2\beta \left[(x + Z)^2 - \gamma^2 \right]_+ \right) \\ &\leq \exp\left(-\frac{1}{2} \left[(x + Z)^2 - \gamma^2 \right]_+ \right) \end{aligned}$$

where the last inequality comes from $\beta \geq 1/4$. We then obtain

$$\begin{aligned} \mathbb{E}\left((x - s_\beta(x + Z)(x + Z))^2 \right) &\leq 2 + 4x^2\mathbb{E}\left(\exp\left(-\frac{1}{2} \left[(x - Z)^2 - \gamma^2 \right]_+ \right) \mathbf{1}_{Z>0} \right) \\ &\leq 2 + 4x^2 \left[\mathbb{P}\left(0 < Z < x/2\right) \exp\left(-\frac{1}{2} \left[(x/2)^2 - \gamma^2 \right] \right) + \mathbb{P}\left(Z > x/2\right) \right] \\ &\leq 2 + 2x^2 \exp\left(-x^2/8 + \gamma^2/2 \right) + 4x^2 \exp\left(-x^2/8 \right). \end{aligned}$$

When $p \geq 3$ and $\beta \leq 1/2$, we have $\gamma \geq \sqrt{2 + 2\log 3}$ and then

$$\sup_{x > 4\gamma} x^2 e^{-x^2/8} = 16\gamma^2 \exp(-2\gamma^2).$$

Therefore,

$$\mathbb{E}\left((x - s_\beta(x + Z)(x + Z))^2 \right) \leq 2 + 16\gamma^2 \left(2e^{-3\gamma^2/2} + 4e^{-2\gamma^2} \right) \leq 0.6\,\gamma^2 \leq c_\beta(p)\gamma^2,$$

where we used again the bound $\gamma^2 \geq 2 + 2\log 3$. The proof of Proposition 2 is complete.

6.5. **Proof of Corollary 1.** We start by proving some results on the approximation of BV functions with the Haar wavelets.

6.5.1. *Approximation of BV functions.* We stick to the setting of Section 4.2 with $0 = x_1 < x_2 < \cdots < x_n < x_{n+1} = 1$ and $n = 2^{J_n}$. For $0 \le j \le J_n$ and $p \in \Lambda(j)$ we define $t_{j,p} = x_{p2^{-j}n+1}$. We also set $\phi_{0,0} = 1$ and

$$\phi_{j,k} = 2^{(j-1)/2} \left(\mathbf{1}_{[t_{j,2k+1}, t_{j,2k+2})} - \mathbf{1}_{[t_{j,2k}, t_{j,2k+1})} \right), \text{ for } 1 \le j \le J_n \text{ and } k \in \Lambda(j).$$

This family of Haar wavelets is orthonormal for the positive semi-definite quadratic form

$$(f,g)_n = \frac{1}{n} \sum_{i=1}^{n} f(x_i) g(x_i)$$

on functions mapping $[0,1]$ into $\mathbb{R}$. For $0 \le J \le J_n$, we write f_J for the projection of f onto the linear space spanned by $\{\phi_{j,k},\ 0 \le j \le J,\ k \in \Lambda(j)\}$ with respect to $(\cdot,\cdot)_n$, namely

$$f_J = \sum_{j=0}^{J} \sum_{k \in \Lambda(j)} c_{j,k} \phi_{j,k}, \quad \text{with } c_{j,k} = (f, \phi_{j,k})_n.$$

We also consider for $1 \le J \le J_n$ an approximation of f *à la* Birgé and Massart [6]

$$\tilde{f}_J = f_{J-1} + \sum_{j=J}^{J_n} \sum_{k \in \Lambda'_J(j)} c_{j,k} \phi_{j,k},$$

where $\Lambda'_J(j) \subset \Lambda(j)$ is the set of indices k we obtain when we select the $K_{j,J}$ largest coefficients $|c_{j,k}|$ amongs $\{|c_{j,k}|,\ k \in \Lambda(j)\}$, with $K_{j,J} = \lfloor (j-J+1)^{-3} 2^{J-2} \rfloor$ for $1 \le J \le j \le J_n$. Note that the number of coefficients $c_{j,k}$ in $\tilde{f}_J$ is bounded from above by

$$1 + \sum_{j=1}^{J-1} 2^{j-1} + \sum_{j \ge J} (j-J+1)^{-3} 2^{J-2} \le 2^{J-1} + 2^{J-2} \sum_{p \ge 1} p^{-3} \le 2^J.$$

Next proposition states approximation bounds for f_J and $\tilde{f}_J$ in terms of the (semi-)norm $\|f\|_n^2 = (f,f)_n$.

Proposition 3. *When f has bounded variation $V(f)$, we have*

$$\|f - f_J\|_n \le 2V(f) 2^{-J/2}, \quad \text{for } J \ge 0 \tag{25}$$

and

$$\|f - \tilde{f}_J\|_n \le c\, V(f) 2^{-J}, \quad \textit{for } J \ge 1 \tag{26}$$

with $c = \sum_{p \ge 1} p^3 2^{-p/2+1}$.

Formulaes (25) and (26) are based on the following fact.

Lemma 2. *When f has bounded variation $V(f)$, we have*

$$\sum_{k \in \Lambda(j)} |c_{j,k}| \le 2^{-(j+1)/2} V(f), \quad \text{for } 1 \le j \le J_n.$$

Proof of the Lemma. We assume for simplicity that f is non-decreasing. Then, we have

$$c_{j,k} =< f, \phi_{j,k} >_n = \frac{2^{(j-1)/2}}{n} \left[\sum_{i \in I^+_{j,k}} f(x_i) - \sum_{i \in I^-_{j,k}} f(x_i) \right],$$

with $I^+_{j,k}$ and $I^-_{j,k}$ defined in Section 4.2. Since $|I^+_{j,k}| = 2^{-j}n$ and f is non-decreasing

$$\begin{aligned} |c_{j,k}| &\leq \frac{2^{(j-1)/2}}{n} |I^+_{j,k}| \left[f(x_{(2k+2)2^{-j}n}) - f(x_{(2k)2^{-j}n}) \right] \\ &\leq 2^{-(j+1)/2} \left[f(x_{(2k+2)2^{-j}n}) - f(x_{(2k)2^{-j}n}) \right], \end{aligned}$$

and Lemma 2 follows. □

We first prove (25). Since the $\{\phi_{j,k},\ k \in \lambda(j)\}$ have disjoint supports, we have for $0 \leq J \leq J_n$

$$\begin{aligned} \|f - f_J\|_n &\leq \sum_{j>J} \| \sum_{k \in \Lambda(j)} c_{j,k} \phi_{j,k} \|_n \\ &\leq \sum_{j>J} \left[\sum_{k \in \Lambda(j)} |c_{j,k}|^2 \underbrace{\|\phi_{j,k}\|^2_n}_{=1} \right]^{1/2} \\ &\leq \sum_{j>J} \sum_{k \in \Lambda(j)} |c_{j,k}|. \end{aligned}$$

Formula (25) then follows from Lemma 2.

To prove (26) we introduce the set $\Lambda''_J(j) = \Lambda(j) \setminus \Lambda'_J(j)$. Then, for $1 \leq J \leq J_n$ we have

$$\begin{aligned} \|f - \tilde{f}_J\|_n &\leq \sum_{j=J}^{J_n} \left[\sum_{k \in \Lambda''_J(j)} |c_{j,k}|^2 \underbrace{\|\phi_{j,k}\|^2_n}_{=1} \right]^{1/2} \\ &\leq \sum_{j=J}^{J_n} \left[\max_{k \in \Lambda''_J(j)} |c_{j,k}| \sum_{k \in \Lambda(j)} |c_{j,k}| \right]^{1/2}. \end{aligned}$$

The choice of $\Lambda'_J(j)$ enforces the inequalities

$$(1 + K_{j,J}) \max_{k \in \Lambda''_J(j)} |c_{j,k}| \leq \sum_{k \in \Lambda''_J(j)} |c_{j,k}| + \sum_{k \in \Lambda'_J(j)} |c_{j,k}| \leq \sum_{k \in \Lambda(j)} |c_{j,k}|.$$

To complete the proof of Proposition 3, we combine this bound with Lemma 2:

$$\begin{aligned} \|f - \tilde{f}_J\|_n &\leq \sum_{j \geq J} 2^{-(j+1)/2} V(f) (1 + K_{j,J})^{-1/2} \\ &\leq \sum_{j \geq J} 2^{-(j+1)/2} V(f) 2^{-(J-2)/2} (j - J + 1)^3. \\ &\leq V(f) 2^{-J} \sum_{p \geq 1} p^3 2^{-p/2+1}. \end{aligned}$$

6.5.2. *Proof of Corollary 1.* First, note that $v_{j,k} = (\phi_{j,k}(x_1), \ldots, \phi_{j,k}(x_n))'$ for $(j,k) \in \Lambda^*$. Then, according to (25) and (26) there exists for any $0 \le J \le J^*$ a model $m \in \mathcal{M}$ fufilling $|m| \le 2^J$ and

$$\begin{aligned} \|\mu - \Pi_{\mathcal{S}_m}\mu\|_n^2 &= \|\mu - \Pi_{\mathcal{S}_*}\mu\|_n^2 + \|\Pi_{\mathcal{S}_*}\mu - \Pi_{\mathcal{S}_m}\mu\|_n^2 \\ &\le 2c^2 V(f)^2 \left(2^{-J^*} \vee 2^{-2J}\right), \end{aligned}$$

with $c = \sum_{p\ge 1} p^3 2^{-p/2+1}$. Putting together this approximation result with Theorem 1 gives

$$\mathbb{E}\left(\|\mu - \hat{\mu}\|_n^2\right) \le C \inf_{0\le J\le J^*} \left[V(f)^2 \left(2^{-J^*} \vee 2^{-2J}\right) + \frac{2^J \log n}{n}\, \sigma^2\right]$$

for some numerical constant C, when

$$\beta \le \frac{1}{4}\phi^{-1}\left(\frac{\log n}{n/2-2}\right).$$

This bound still holds true when

$$\frac{1}{4}\phi^{-1}\left(\frac{\log n}{n/2-2}\right) \le \beta \le \frac{1}{2}\phi^{-1}\left(\frac{\log(n/2)}{n/2}\right),$$

see Proposition 4 in the Appendix. To conclude the proof of Corollary 1, we apply the previous bound with J given by the minimum between J^* and the smallest integer such that

$$2^J \ge V(f)^{2/3}\left(\frac{n}{\sigma^2 \log n}\right)^{1/3}.$$

6.6. **Proof of Corollary 2.** First, according to the inequality $\binom{n}{k} \le (en/k)^k$ we have the bound for $m \in \mathcal{M}_J$

$$\begin{aligned} -\log \pi_m &\le \log 2^J + \sum_{j=J}^{J_*} \frac{2^J}{(j-J+1)^3}\, \log\left(e 2^{j-J+1}(j-J+1)^3\right) \\ &\le 2^J \left(1 + \sum_{k\ge 1} k^{-3}\left(1 + 3\log k + k\log 2\right)\right) \\ &\le 4 \cdot 2^J. \end{aligned} \tag{27}$$

Second, when f belongs to some Besov ball $\mathcal{B}^\alpha_{p,\infty}(R)$ with $1/p < \alpha < r$, Birgé and Massart [6] gives the following approximation results. There exists a constant $C > 0$, such that for any $J \le J_*$ and $f \in \mathcal{B}^\alpha_{p,\infty}(R)$, there exists $m \in \mathcal{M}_J$ fulfilling

$$\|f - \bar{\Pi}_{\mathcal{F}_m} f\|_\infty \le \|f - \bar{\Pi}_{\mathcal{F}_*} f\|_\infty + \|\bar{\Pi}_{\mathcal{F}_*} f - \bar{\Pi}_{\mathcal{F}_m} f\|_\infty \le C \max\left(2^{-J_*(\alpha-1/p)}, 2^{-\alpha J}\right)$$

where $\bar{\Pi}_{\mathcal{F}}$ denotes the orthogonal projector onto $\mathcal{F}$ in $L^2([0,1])$. In particular, under the previous assumptions we have

$$\begin{aligned} \|\mu - \Pi_{\mathcal{S}_m}\mu\|_n^2 &\leq \frac{1}{n}\sum_{i=1}^{n}\left[f(x_i) - \bar{\Pi}_{\mathcal{F}_m} f(x_i)\right]^2 \\ &\leq \|f - \bar{\Pi}_{\mathcal{F}_m} f\|_\infty^2 \leq C^2 \max\left(2^{-2\alpha J}, 2^{-2J_*(\alpha - 1/p)}\right), \end{aligned} \tag{28}$$

To conclude the proof of Corollary 2, we combine Theorem 1 together with (21), (27) and (28) for

$$J = \min\left(J_*\ ,\ \left\lfloor \frac{\log[\max(n/\sigma^2, 1)]}{(2\alpha+1)\log 2} \right\rfloor + 1\right).$$

References

[1] Akaike, H. (1969) *Statistical predictor identification.* Annals of the Institute of Statistical Mathematics. **22**, 203–217

[2] Baraud, Y., Giraud, C. and Huet, S. (2006) *Gaussian model selection with unknown variance.* http://arXiv:math/0701250v1

[3] Barron, A. (1987) *Are Bayesian rules consistent in information?* Open problems in Communication and Computation. T. Cover and B. Gopinath, Eds. Springer-Verlag.

[4] Barron, A., Birgé, L. and Massart, P. (1999) *Risk bounds for model selection via penalization.* Probab. Theory Related Fields **113** no. 3, 301–413.

[5] Barron, A. and Cover, T. (1991) *Minimum complexity density estimation.* IEEE Trans. Inform. Theory **37** no. 4, 1034–1054.

[6] Birgé, L. and Massart, P. (2000) *An adaptive compression algorithm in Besov spaces.* Constr. Approx. **16**, no. 1, 1–36.

[7] Birgé, L. and Massart, P. (2001) *Gaussian model selection.* J. Eur. Math. Soc. (JEMS) **3**, no. 3, 203–268.

[8] Birgé, L. and Massart, P. (2006) *A generalized C_p criterion for Gaussian model selection.* To appear in Probab. Theory Related Fields.

[9] Bunea, F., Tsybakov, A. and Wegkamp, M. (2007) *Aggregation for Gaussian regression.* Ann. Statist. **35**, no. 4, 1674–1697.

[10] Catoni, O. (1997) *Mixture approach to universal model selection.* Laboratoire de l'Ecole Normale Supérieure, Paris. Preprint 30.

[11] Catoni, O. (1999) *Universal aggregation rules with exact bias bounds.* Laboratoire de Probabilités et Modèles Aléatoires, CNRS, Paris. Preprint 510.

[12] Devore, R. and Lorentz, G. (1993) *Constructive approximation.* Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], **303**. Springer-Verlag

[13] Donoho, D. and Johnstone, I. (1994) *Ideal spatial adaptation by wavelet shrinkage.* Biometrika **81**, no. 3, 425–455.

[14] Hall, P., Kay, J. and Titterington, D. M. (1990). *Asymptotically optimal differencebased estimation of variance in nonparametric regression.* Biometrika **77**, 521–528.

[15] Hartigan, J.A. (2002) *Bayesian regression using Akaike priors.* Yale University, New Haven. Preprint.

[16] Lenth, R. (1989) *Quick and easy analysis of unreplicated factorials.* Technometrics **31**, no. 4, 469–473.

[17] Leung, G. and Barron, A. (2006) *Information Theory and Mixing Least-Squares Regressions.* IEEE Transact. Inf. Theory **52**, no. 8, 3396–3410.

[18] Mallows, C. (1973) *Some comments on C_p.* Technometrics **15**, 661-675.

[19] Munk, A., Bissantz, N., Wagner, T. and Freitag, G. (2005). *On difference based variance estimation in nonparametric regression when the covariate is high dimensional.* J. Roy. Statist. Soc. B **67**, 19–41.

[20] Rice, J. (1984). *Bandwidth choice for nonparametric kernel regression.* Ann. Statist. **12**, 1215–1230.

[21] Stein, C. (1981) *Estimation of the mean of a multivariate normal distribution.* Ann. Statist. **9**, 1135–1151.

[22] Tong, T., Wang, Y. (2005) *Estimating residual variance in nonparametric regression using least squares.* Biometrika **92**, no. 4, 821–830.
[23] Tsybakov, A. (2003) *Optimal rates of aggregation.* COLT-2003, Lecture Notes in Artificial Intelligence, **2777**, Springer, Heidelberg, 303–313.
[24] Yang, Y. and Barron, A. (1999) *Information-theoretic determination of minimax rates of convergence* Ann. Statist. **27**, no. 5, 1564–1599.
[25] Yang, Y. (2000) *Combining different procedures for adaptive regression.* J. Multivariate Anal. **74**, no. 1, 135–161.
[26] Yang, Y. (2000) *Mixing Strategies for Density Estimation.* Ann. Statist. **28**, 75–87.
[27] Yang, Y. (2003) *Regression with multiple candidate models: selecting or mixing?* Statist. Sinica **13**, no. 3, 783–809.
[28] Yang, Y. (2004) *Combining forecasting procedures: some theoretical results.* Econometric Theory **20**, no. 1, 176–222
[29] Wang, L., Brown, L., Cai, T. and Levine, M. *Effect of mean on variance function estimation in nonparametric regression.* To appear in Ann. Statist.

APPENDIX

The Appendix is devoted to the proof of the bound $c_\beta(p) \leq 16$ and gives further explanations on the Remark 4, Section 3.2. The results are stated in Appendix A and the proofs (of minor interest) can be found in Appendix B.

Appendix A. Two bounds

A.1. **When the variance is known.** In this section, we assume that the noise level σ is known. We remind that $3 \leq p \leq n$ and $\{v_1, \dots, v_p\}$ is an orthonormal family of vectors in $\mathbb{R}^n$. Next lemma gives an upper bound on the Euclidean risk of the estimator

$$\hat{\mu} = \sum_{j=1}^{p} \left(\frac{e^{\beta Z_j^2/\sigma^2}}{p e^{\beta\lambda} + e^{\beta Z_j^2/\sigma^2}} Z_j \right) v_j, \qquad \text{with } \lambda \geq 2 \text{ and } Z_j =< Y, v_j >,\ j = 1, \dots, p.$$

Lemma 3. *For any $\lambda \geq 2$ and $\beta > 0$, we set $\gamma^2 = \lambda + \beta^{-1} \log p$.*

1. *For $1/4 \leq \beta \leq 1/2$ and $a \in \mathbb{R}$ we have the bound*

$$\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)} (a+\varepsilon) \right)^2 \right] \leq 16 \left[\min\left(a^2, \gamma^2\right) + \gamma^2/p \right], \tag{29}$$

   where ε is distributed as a standard Gaussian random variable.
2. *As a consequence, for any $0 < \beta \leq 1/2$ and $\lambda \geq 2$ the risk of the estimator $\hat{\mu}$ is upper bounded by*

$$\mathbb{E}\left(\|\mu - \hat{\mu}\|^2 \right) \leq 16 \inf_{m \in \mathcal{M}} \left[\|\mu - \mu_m\|^2 + \gamma^2 |m| \sigma^2 + \gamma^2 \sigma^2 \right]. \tag{30}$$

Note that (29) enforces the bound $c_\beta(p) \leq 16$. This constant 16 is certainly far from optimal. Indeed, when $\beta = 1/2$, $\lambda = 2$ and $3 \leq p \leq 10^6$, Figure 1 in Section 3.2 shows that the bounds (29) and (30) hold with 16 replaced by 1.

A.2. **When the variance is unknown.** We consider the same framework, except that the noise level σ is not known. Next proposition provides a risk bound for the estimator

$$\hat{\mu} = \sum_{j=1}^{p} (c_j Z_j) v_j, \quad \text{with} \quad Z_j =< Y, v_j > \text{and} \quad c_j = \frac{\exp\left(\beta Z_j^2/\hat{\sigma}^2\right)}{p\exp\left(b\right) + \exp\left(\beta Z_j^2/\hat{\sigma}^2\right)}. \tag{31}$$

We remind that $\phi(x) = (x - 1 - \log x)/2$.

Proposition 4. *Assume that β and p fulfill the conditions,*

$$p \geq 3, \quad 0 < \beta < 1/2 \quad \text{and} \quad p + \frac{\log p}{\phi(2\beta)} \leq n. \tag{32}$$

Assume also that b is not smaller than 1. Then, we have the following upper bound on the L^2-risk of the estimator $\hat{\mu}$ defined by (31)

$$\mathbb{E}\left(\|\mu - \hat{\mu}\|^2\right) \leq 16 \inf_{m \in \mathcal{M}} \left[\|\mu - \mu_m\|^2 + \frac{1}{\beta}\left(b + \log p\right)(|m|+1)\bar{\sigma}^2 + (2 + b + \log p)\sigma^2 \right], \tag{33}$$

where $\bar{\sigma}^2 = \sigma^2 + \|\mu - \Pi_{\mathcal{S}_*}\mu\|^2/(n-p)$.

We postpone the proof of Proposition 4 to Section B.2.

Appendix B. Proofs

B.1. **Proof of Lemma 3.** When $\beta \leq 1/4$ the bound (30) follows from a slight variation of Theorem 5 in [17]. When $1/4 < \beta \leq 1/2$, it follows from (29) (after a rescaling by σ^2 and a summation). So all we need is to prove (29). For symmetry reason, we restrict to the case $a > 0$.

To prove Inequality (29), we first note that

$$\begin{aligned}
&\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)}(a+\varepsilon)\right)^2\right] \\
&\quad = \mathbb{E}\left[\left(\frac{a}{1 + p^{-1}\exp\left(\beta\left[(a+\varepsilon)^2 - \lambda\right]\right)} - \frac{\varepsilon}{1 + p\exp(\beta\left[\lambda - (a+\varepsilon)^2\right])}\right)^2\right] \\
&\quad \leq 2a^2\,\mathbb{E}\left[\left(\frac{1}{1 + p^{-1}\exp\left(\beta\left[(a+\varepsilon)^2 - \lambda\right]\right)}\right)^2\right] + 2\mathbb{E}\left[\left(\frac{\varepsilon}{1 + p\exp(\beta\left[\lambda - (a+\varepsilon)^2\right])}\right)^2\right]. \qquad (34)
\end{aligned}$$

and then investigate apart the four cases $0 \leq a \leq 2\gamma^{-1}$, $2\gamma^{-1} \leq a \leq 1$, $1 \leq a \leq \sqrt{3}\,\gamma$ and $a \geq \sqrt{3}\,\gamma$. Inequality (29) will follow from Inequalities (36), (37), (38) and (39).

Case $0 \leq a \leq 2\gamma^{-1}$. From (34) we get

$$\begin{aligned}
\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)}(a+\varepsilon)\right)^2\right] &\leq 2a^2 + 2\mathbb{E}\left[\left(\frac{\varepsilon}{1 + p\exp(\beta\left[\lambda - (a+\varepsilon)^2\right])}\right)^2\right] \\
&\leq 2a^2 + 4\mathbb{E}\left[\left(\frac{\varepsilon\mathbf{1}_{\varepsilon>0}}{1 + p\exp(\beta\left[\lambda - (a+\varepsilon)^2\right])}\right)^2\right] \\
&\leq 2a^2 + 4\mathbb{E}\left[\varepsilon^2\mathbf{1}_{\varepsilon>0}\exp\left(-2\beta\left[\gamma^2 - (a+\varepsilon)^2\right]_+\right)\right]
\end{aligned}$$

with $\gamma^2 = \lambda + \beta^{-1}\log p$. Expanding the expectation gives

$$\begin{aligned}
&\mathbb{E}\left[\varepsilon^2 \mathbf{1}_{\varepsilon>0} \exp\left(-2\beta\left[\gamma^2 - (a+\varepsilon)^2\right]_+\right)\right] \\
&\quad = \quad e^{-2\beta\gamma^2}\int_0^{\gamma-a} x^2 e^{2\beta(a+x)^2 - x^2/2}\,\frac{dx}{\sqrt{2\pi}} + \int_{\gamma-a}^{\infty} x^2 e^{-x^2/2}\,\frac{dx}{\sqrt{2\pi}}
\end{aligned}$$

Note that when $p \geq 3$, $\lambda \geq 2$ and $\beta \leq 1/2$, we have $\gamma \geq 2$. Therefore, when $0 \leq a \leq 1$, an integration by parts in the second integral gives

$$\begin{aligned}
\int_{\gamma-a}^{\infty} x^2 e^{-x^2/2}\,\frac{dx}{\sqrt{2\pi}} &= \left[\frac{-xe^{-x^2/2}}{\sqrt{2\pi}}\right]_{\gamma-a}^{\infty} + \int_{\gamma-a}^{\infty} e^{-x^2/2}\,\frac{dx}{\sqrt{2\pi}} \\
&\leq \frac{2(\gamma-a)}{\sqrt{2\pi}}\, e^{-(\gamma-a)^2/2}.
\end{aligned}$$

For the first integral, since $\beta > 1/4$, we have the bound

$$e^{-2\beta\gamma^2}\int_0^{\gamma-a} x^2 e^{2\beta(a+x)^2 - x^2/2}\,\frac{dx}{\sqrt{2\pi}} \leq e^{-(\gamma-a)^2/2}\int_0^{\gamma-a} x^2\,\frac{dx}{\sqrt{2\pi}} \leq \frac{(\gamma-a)^3}{3\sqrt{2\pi}}\, e^{-(\gamma-a)^2/2}.$$

Besides an integration by parts also gives

$$\begin{aligned}
&e^{-2\beta\gamma^2}\int_0^{\gamma-a} x^2 e^{2\beta(a+x)^2 - x^2/2}\,\frac{dx}{\sqrt{2\pi}} \\
&\leq \quad (4\beta-1)^{-1} e^{-2\beta\gamma^2}\int_0^{\gamma-a} (4\beta(a+x) - x) x e^{2\beta(a+x)^2 - x^2/2}\,\frac{dx}{\sqrt{2\pi}} \\
&\leq \quad \frac{e^{-2\beta\gamma^2}}{(4\beta-1)\sqrt{2\pi}}\left[x e^{2\beta(a+x)^2 - x^2/2}\right]_0^{\gamma-a} = \frac{\gamma-a}{(4\beta-1)\sqrt{2\pi}}\, e^{-(\gamma-a)^2/2}.
\end{aligned}$$

Putting pieces together, we obtain for $0 \leq a \leq 1$ and $1/4 < \beta \leq 1/2$

$$\begin{aligned}
&\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)}(a+\varepsilon)\right)^2\right] \\
&\leq \quad 2a^2 + \frac{8 + 4\min\left((4\beta-1)^{-1}, 3^{-1}(\gamma-a)^2\right)}{\sqrt{2\pi}}(\gamma-a)e^{-(\gamma-a)^2/2} \\
&\leq \quad 2a^2 + \frac{8 + 4\min\left((4\beta-1)^{-1}, 3^{-1}(\gamma-a)^2 e^{-(1/2-\beta)(\gamma-a)^2}\right)}{\sqrt{2\pi}}(\gamma-a)e^{-\beta(\gamma-a)^2} \\
&\leq \quad 2a^2 + \frac{8 + 4\min\left((4\beta-1)^{-1}, [3e(1/2-\beta)]^{-1}\right)}{\sqrt{2\pi}}\,\gamma e^{-\beta(\gamma-a)^2} \\
&\leq \quad 2a^2 + 5.6\,\gamma\, e^{-\beta(\gamma-a)^2}. \qquad (35)
\end{aligned}$$

Furthermore, when $0 \le a \le 2\gamma^{-1}$, we have $(\gamma - a)^2 \ge \gamma^2 - 4$. The inequalities (35) and $\lambda \ge 2$ thus give

$$\begin{aligned}\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)}(a+\varepsilon)\right)^2\right] &\le 2a^2 + 5.6e^{(4-\lambda)\beta}\gamma/p \\ &\le 2a^2 + 8\,\gamma^2/p\,. \end{aligned} \tag{36}$$

Case $2\gamma^{-1} \le a \le 1$. Starting from (35) we have for $2\gamma^{-1} \le a \le 1$

$$\begin{aligned}\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)}(a+\varepsilon)\right)^2\right] &\le 2a^2 + \frac{5.6\,\gamma^3 e^{-\beta(\gamma-1)^2}}{\gamma^2} \\ &\le 2a^2 + \frac{5.6\,\gamma^3 e^{-(\gamma-1)^2/4}}{\gamma^2} \\ &\le 2a^2 + 56\gamma^{-2} \;\le\; 16a^2. \end{aligned} \tag{37}$$

Case $1 \le a \le \sqrt{3}\,\gamma$. From (34) we get

$$\begin{aligned}\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)}(a+\varepsilon)\right)^2\right] &\le 2(a^2+1) \\ &\le 4a^2 \;\le\; 12\min(a^2,\gamma^2). \end{aligned} \tag{38}$$

Case $a > \sqrt{3}\gamma$. From (34) we get

$$\begin{aligned}&\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)}(a+\varepsilon)\right)^2\right] \\ &\le 4a^2\,\mathbb{E}\left[\left(\frac{1}{1 + p^{-1}\exp\left(\beta\left[(a-\varepsilon)^2 - \lambda\right]\right)}\right)^2 \mathbf{1}_{\varepsilon>0}\right] + 2 \\ &\le 4a^2\,\mathbb{E}\left[\exp\left(-2\beta\left[(a-\varepsilon)^2 - \gamma^2\right]_+\right)\mathbf{1}_{\varepsilon>0}\right] + 2 \\ &\le 2a^2\exp\left(-2\beta\left[(2a/3)^2 - \gamma^2\right]_+\right) + 4a^2\mathbb{P}\left(\varepsilon > a/3\right) + 2.\end{aligned}$$

Since $(2a/3)^2 > a^2/9 + \gamma^2$, we finally obtain

$$\begin{aligned}&\mathbb{E}\left[\left(a - \frac{\exp\left(\beta(a+\varepsilon)^2\right)}{p\exp(\beta\lambda) + \exp\left(\beta(a+\varepsilon)^2\right)}(a+\varepsilon)\right)^2\right] \\ &\le 2a^2\exp\left(-2\beta a^2/9\right) + 4a^2\exp\left(-a^2/18\right) + 2 \\ &\le 42 \le 11\gamma^2.\end{aligned} \tag{39}$$

B.2. **Proof of Proposition 4.** We can express the weights c_j appearing in (31) in the following way

$$c_j = \frac{e^{\hat\beta Z_j^2/\sigma^2}}{pe^{\hat\beta\hat\lambda} + e^{\hat\beta Z_j^2/\sigma^2}}$$

with $\hat\beta = \beta\sigma^2/\hat\sigma^2$ and $\hat\lambda = b/\hat\beta$. Note that $\hat\beta \le 1/2$ enforces $\hat\lambda \ge 2$ since $b \ge 1$.

Since $\hat\sigma^2$ is independent of the Z_js, we can work conditionally on $\hat\sigma^2$. When $\hat\beta$ is not larger than $1/2$ we apply Lemma 3 and get

$$\begin{aligned}\mathbb{E}\left(\|\mu-\hat\mu\|^2 \,|\, \hat\sigma^2\right)\mathbf{1}_{\{\hat\beta\le 1/2\}} &\le 16\inf_{m\in\mathcal{M}}\left[\|\mu-\mu_m\|^2 + \left(\hat\lambda + \hat\beta^{-1}\log p\right)(|m|+1)\sigma^2\right]\mathbf{1}_{\{\hat\beta\le 1/2\}} \\ &\le 16\inf_{m\in\mathcal{M}}\left[\|\mu-\mu_m\|^2 + \frac{1}{\beta}\left(b+\log p\right)(|m|+1)\hat\sigma^2\right]\mathbf{1}_{\{\hat\beta\le 1/2\}}. \qquad (40)\end{aligned}$$

When $\hat\beta$ is larger than $1/2$, we use the following bound.

Lemma 4. *Write Z for a Gaussian random variable with mean a and variance σ^2. Then for any $\hat\lambda \ge 0$ and $\hat\beta > 1/2$*

$$\mathbb{E}\left[\left(a - \frac{e^{\hat\beta Z^2/\sigma^2}}{pe^{\hat\beta\hat\lambda} + e^{\hat\beta Z^2/\sigma^2}}\,Z\right)^2\right] \le 6\left(\hat\lambda + \hat\beta^{-1}\log p\right)\sigma^2 + 36\sigma^2. \qquad (41)$$

Proof. First, we write ε for a standard Gaussian random variable and obtain

$$\begin{aligned}&\mathbb{E}\left[\left(a - \frac{e^{\hat\beta Z^2/\sigma^2}}{pe^{\hat\beta\hat\lambda} + e^{\hat\beta Z^2/\sigma^2}}\,Z\right)^2\right] \\ &\quad= \mathbb{E}\left[\left(\frac{a}{1+p^{-1}\exp\left(\hat\beta\left[(a/\sigma+\varepsilon)^2-\hat\lambda\right]\right)} - \frac{\sigma\varepsilon}{1+p\exp\left(\hat\beta\left[\hat\lambda-(a/\sigma+\varepsilon)^2\right]\right)}\right)^2\right] \\ &\quad\le 2(a^2+\sigma^2).\end{aligned}$$

Whenever a^2 is smaller than $3\left(\hat\lambda + \hat\beta^{-1}\log p\right)\sigma^2$, this quantity remains smaller than $6\left(\hat\lambda + \hat\beta^{-1}\log p\right)\sigma^2 + 2\sigma^2$.

When a^2 is larger than $3\left(\hat\lambda + \hat\beta^{-1}\log p\right)\sigma^2$ we follow the same lines as in the last case of Section B.1 and get

$$\begin{aligned}&\mathbb{E}\left[\left(a - \frac{e^{\hat\beta Z^2/\sigma^2}}{pe^{\hat\beta\hat\lambda} + e^{\hat\beta Z^2/\sigma^2}}\,Z\right)^2\right] \\ &\qquad\le 2\sigma^2 + 4a^2\mathbb{P}\left(\varepsilon > |a|/(3\sigma)\right) + 2a^2p^2\exp\left[-2\hat\beta\left(\left(\frac{2a}{3\sigma}\right)^2 - \hat\lambda\right)\right].\end{aligned}$$

Since $(2a/3)^2 \geq a^2/9 + \left(\hat{\lambda} + \hat{\beta}^{-1}\log p\right)\sigma^2$, we obtain

$$\begin{aligned}\mathbb{E}\left[\left(a - \frac{e^{\hat{\beta}Z^2/\sigma^2}}{pe^{\hat{\beta}\hat{\lambda}} + e^{\hat{\beta}Z^2/\sigma^2}}\, Z\right)^2\right] &\leq 2\sigma^2 + 4a^2\exp\left(-a^2/(18\sigma^2)\right) + 2a^2\exp\left(-a^2/(9\sigma^2)\right)\\ &\leq 36\sigma^2.\end{aligned}$$

□

From (41), we obtain after summation

$$(42)\quad \mathbb{E}\left(\|\mu-\hat{\mu}\|^2 \,|\, \hat{\sigma}^2\right)\mathbf{1}_{\left\{\hat{\beta}>1/2\right\}} \leq p\left[\frac{6}{\beta}\left(b+\log p\right)\hat{\sigma}^2 + 36\sigma^2\right]\mathbf{1}_{\left\{\hat{\beta}>1/2\right\}} + \|\mu - \Pi_{\mathcal{S}_*}\mu\|^2\mathbf{1}_{\left\{\hat{\beta}>1/2\right\}}.$$

Futhermore $\hat{\sigma}^2$ is smaller than $2\beta\sigma^2$ when $\hat{\beta}$ is larger than $1/2$, so taking the expectation of (42) and (40) gives

$$\begin{aligned}\mathbb{E}\left(\|\mu-\hat{\mu}\|^2\right) &\leq 16\inf_{m\in\mathcal{M}}\left[\|\mu-\mu_m\|^2 + \frac{1}{\beta}\left(b+\log p\right)\left(|m|+1\right)\bar{\sigma}^2\right]\\ &+ p\left[12\left(b+\log p\right)+36\right]\mathbb{P}\left(\hat{\sigma}^2 < 2\beta\sigma^2\right)\sigma^2.\end{aligned}$$

The random variable $\hat{\sigma}^2/\sigma^2$ is stochastically larger than a random variable X distributed as a χ^2 of dimension $n-p$ divided by $n-p$. The Lemma 1 gives

$$\mathbb{P}\left(X \leq 2\beta\right) \leq (2\beta)^{N/2}\exp\left(\frac{N}{2}(1-2\beta)\right) \leq \exp\left(-N\phi(2\beta)\right)$$

and then $\mathbb{P}\left(\hat{\sigma}^2 < 2\beta\sigma^2\right) \leq \exp(-(n-p)\phi(2\beta))$, so Condition (32) ensures that

$$p\,\mathbb{P}\left(\hat{\sigma}^2 < 2\beta\sigma^2\right) \leq 1.$$

Finally, since $12(b+\log p)+36$ is smaller than $16(2+b+\log p)$ we get (33).

Université de Nice Sophia-Antipolis, Laboratoire J-A Dieudonné, Parc Valrose, 06108 Nice cedex 02

E-mail address: cgiraud@math.unice.fr